GHENT UNIVERSITY

FACULTY OF SCIENCES

Retention and integration of gene duplicates in eukaryotes

Jonas Defoort

Ghent University
Faculty of Sciences
Department of Plant Biotechnology and Bioinformatics
VIB center of plant systems biology

Dissertation submitted in fulfilment of the requirements for the degree of Doctor of Philosophy (PhD).

Academic year 2017-2018

EXAMINATION COMMISION

- **Prof. Dr. Geert De Jaeger (Chair)**
  - Department of Plant Biotechnology and Bioinformatics (WE09), Ghent University; VIB Center for Plant Systems Biology
- **Prof. Dr. Yves van de Peer (Promotor)**
  - Department of Plant Biotechnology and Bioinformatics (WE09), Ghent University; VIB Center for Plant Systems Biology
- **Prof. Dr. Klaas Vandepoele**
  - Department of Plant Biotechnology and Bioinformatics (WE09), Ghent University; VIB Center for Plant Systems Biology
- **Prof. Dr. Pieter De Bleser**
  - Data mining and modeling for biomedicine - Bioinformatics Core, Inflammation Research Center (IRC) VIB - Ghent University
- **Dr. Eshchar Mizrachi**
  - University of Pretoria, Department of Genetics, Forestry and Agricultural Biotechnology Institute.
  - University of Pretoria, Centre for Bioinformatics and Computational Biology, Genomics Research Institute.
- **Dr. Pieter Audenaert**
  - Universiteit Gent - imec
- **Dr. Lorenzo Carretero Paulet**
  - Department of Plant Biotechnology and Bioinformatics (WE09), Ghent University; VIB Center for Plant Systems Biology
- **Dr. Oren Tzfadia**
  - VIB Bioinformatics Core

ACKNOWLEDGEMENTS

My four-year journey has come to an end. This journey wouldn't have been possible without the contribution of a lot of people. Now that I'm there the only thing that is left is to thank everyone who contributed along the past four years.

To begin, I would like to thank the members of the jury for carefully revising my work, the nice discussions and the effort you put into it. Your contributions really raised my thesis to a higher level.

Thank you, prof. dr. Yves Van de Peer, for allowing me to start this adventure and the support which allowed me to perform my PhD.

Thank you, biocomp and biocomp-alumni, for joining me the past four fantastic years. It was a pleasure and honour to work with all of you. I wish you all the best in the future. Keep on rocking in the science world! Especially, I would like to thank Vanessa and Riet. You've supported me from the beginning, pushed me in the right direction and stood by my side along the way. Also, I would like to thank Michiel a.k.a. 'Captain Awkward'. It would not have been the same without you and please keep in mind 'Watch out for those yellow blocks!'.

Mijn doctoraat was natuurlijk ook niet mogelijk geweest zonder de nodige steun en ontspanning naast het werk. Ik wil graag vrienden en familie bedanken voor alle fijne momenten die we afgelopen jaren samen beleefd hebben.

Bedankt 'Bende van Bart' voor de geestelijk verruimende momenten, het filosofisch gepalaver en de dagen waarop jullie mijn productiviteit om zeep hielpen.

Bedankt 'Gentation' om de week te breken, de week te eindigen en de onvergetelijke avonturen die we samen beleefd hebben. Hierbij moet ik ook RDS bedanken om mijn leven te sponseren met onvergetelijke ervaringen. It's the journey, not the destination!

Bedankt mama, papa, Ruben en Lieselot! Jullie hebben mij gemaakt tot wie ik ben en de fundamenten gelegd voor waar ik nu sta. Ik wil jullie dan ook van harte bedanken voor alles wat jullie gedaan hebben. Daarnaast wil ik ook een klein wonder genaamd 'Nel' bedanken. Jouw schattige lach brengt het beste in mensen naar boven!

Onder het motto "save the best for last": Bedankt Justine! Dit blokje is te kort om jou te bedanken. Maar ik wil dat je weet wat je voor mij betekend. Je hielp mij doorheen moeilijkere en stressvolle momenten, bracht rust in mijn hoofd en leerde me ook te genieten van de kleine dingen des levens.

Tot slot wil ik iedereen bedanken die aanwezig is vandaag. Zo hebben jullie de finale bijdrage geleverd aan dit doctoraat.

Thank everyone who's present here today. With this final contribution you make this moment a nice ending of my PhD!

Van harte bedankt – thank you very much!

Jonas

SUMMARY

The duplication of genes and whole genomes is an important mechanism to increase genomic novelty. In plants, paleopolyploidy events (ancient whole genome duplication events) are found at the basis of all important plant lineages (e.g. at the basis of angiosperms, eudicots and monocots) and within the domesticated crops polyploids have been selected for their higher yield and better fruits. Therefore, there is a need to study how genomes change after duplication events, how the resulting duplicates evolve over time, which molecular mechanisms influence duplicate loss and retention, and how these duplicates are integrated into the existing gene network context.

Gene duplicate loss and retention in flowering plants

Gene duplicates, either generated through whole genome duplications (WGD) or small-scale genome duplications (SSD), are believed to play an important role in generating evolutionary novelty and adaptation. Hence, “Which genes undergo duplication and are preserved following duplication?” and “Why are certain duplicates longer retained?” are important questions. It has been observed that gene duplicability, or the ability of genes to be retained following duplication, is a non-random process. For example, certain biological function categories tend to be preferentially duplicated through WGD, while other functions are enriched among SSD duplicates. However, an overarching view of ‘gene duplicability’ is lacking and the mechanisms influencing loss and retention of gene duplicates over evolutionary time are not yet fully elucidated.

In chapter 2, we present a large-scale study in which we investigated duplicate retention for gene families shared between 37 flowering plant species (angiosperm core gene families). For most gene families, we observe a strikingly consistent pattern of gene duplicability across species, with gene families being either primarily single-copy or multi-copy in all species. An intermediate class contains gene families that are often retained in duplicate for periods extending to tens of millions of years after whole-genome duplication, but ultimately appear to be largely restored to singleton status, suggesting that these genes may be dosage balance-sensitive. The distinction between single-copy and multi-copy gene families is reflected in their functional annotation, with single-copy genes being mainly involved in the maintenance of genome stability and organelle function and multi-copy genes in signalling, transport and metabolism. The intermediate class was overrepresented in regulatory genes, further suggesting that these represent putative dosage-balance sensitive genes.

In Chapter 3, we investigated the impact of protein-protein interactions (PPI) on the evolutionary and functional fate of WGD and SSD duplicates in Arabidopsis, tomato and maize. Using a duplicate classification based on gene family, phylogenetic trees and synteny analyses, a large RNAseq expression compendium, and an extensive protein interaction network, significant divergence at the level of sequence, expression pattern and protein interaction partners could be observed between tandem (SSD) and block (WGD) duplicates. Furthermore, consistently, duplicates involved in PPIs tend to be more evolutionary constrained than their counterparts without interactions in Arabidopsis, tomato and maize. Duplicates with PPI can i) contribute to explain the loss and retention patterns of gene families across angiosperms, and ii) are enriched in gene families predicted to be dosage balance sensitive based on their reciprocal gene retention pattern. In

summary chapter 4 shows based on sequence, expression, interaction and gene retention data the influence of PPIs on the evolution of gene duplicates and that duplication mode and PPIs can be considered as characteristics of a gene family influencing retention of duplicates throughout all angiosperms.

TranSeq: high-throughput 3'-end sequencing

In chapter 4, we present TranSeq, a high-throughput 3'-end sequencing procedure. TranSeq requires 10- to 20-fold fewer sequence reads than the current transcriptomics procedures. TranSeq significantly reduces costs and allows a great increase in size of sample sets analysed in a single experiment. Furthermore, mapping TranSeq reads to the reference tomato genome facilitated the annotation of new transcripts improving > 45% of the existing gene models. Hence, TranSeq is anticipated to boost large-scale transcriptome assays and impact the spatial and temporal resolution of gene expression data and their visualization, in both model and non-model plant species.

Integrated network motif modules

In chapter 5, we present conceptual insights in the topology of eukaryotic integrated gene regulatory networks in a functional, dynamic and evolutionary context. Different types of molecular interactions closely work together in these networks to establish proper gene expression in time and space, but many questions remain on how they specifically influence one another and how they together coordinate gene regulation, especially in higher eukaryotes. To get a systems level understanding of how different molecular interactions interrelate to form a coordinated response in gene regulation, we developed a framework to construct and investigate integrated gene regulatory networks consisting of undirected protein-protein, genetic and homologous interactions, and directed interactions of protein-DNA, regulatory and miRNA-mRNA interactions in the worm *Caenorhabditis elegans* and the plant *Arabidopsis thaliana*. Specifically, we look at network motifs and their clustered modules. We found that composite network motifs cluster together into biologically relevant network modules in integrated gene regulatory networks of worm and plant, thereby relating network topology to function. We integrated gene expression profiles to obtain dynamic network motif modules. We also dissected networks and modules via phylogenetic decomposition to associate the evolutionary age of genes with topological and functional properties. Moreover, we discuss the impact of specific data types and incomplete networks.

SAMENVATTING

Genduplicatie is een belangrijk proces voor het introduceren van nieuwe kenmerken en wijzigingen in planten. In alle commerciële gewassen zijn oude volledige genoom verdubbelingen gedetecteerd (bijv. aan de basis van alle bloemplanten). In gedomesticeerde gewassen zijn polyploïde planten geselecteerd vanwege hogere opbrengst en/of interessante vruchteigenschappen. Het is dan ook cruciaal om te bestuderen hoe genomen veranderen na duplicatie, hoe deze genduplicaten evolueren en hoe deze duplicaten geïntegreerd worden in het reeds bestaande netwerk.

Verlies en behoud van gen duplicaten in bloemplanten.

Genduplicaten kunnen gecreëerd worden door het volledige genoom te dupliceren of door een kleinschalige duplicatie. Het is echter zo dat na duplicatie de meeste genen verloren gaan, maar het behoud van bepaalde duplicaten en het duplicatie mechanisme zijn echter geen willekeurige processen. Sommige functionele categorieën dupliceren bij voorkeur door volledige genoom duplicatie, terwijl andere kleinschalige duplicaties verkiezen. Ondanks vele studies in dit veld ontbrak er nog een algemeen beeld van de dupliceerbaarheid van genen en zijn de mechanismen die het verlies en behoud van duplicaten beïnvloeden onderbelicht.

Hiervoor gaan we in hoofdstuk twee dieper in op de vragen “Welke genen dupliceren?”, ”Welke duplicaten blijven behouden?” en “Waarom blijven bepaalde genen langer behouden?”. In dit hoofdstuk onderzoeken we op grote schaal het behoud van genduplicaten voor genfamilies die genen bevatten uit 37 verschillende bloemplanten. Voor de meeste genfamilies werd een consistent patroon van genduplicatie in alle species geobserveerd. De grootste groep van genfamilies heeft ofwel preferentieel één gen in elk species, of heeft preferentieel meerdere kopieën in alle species. In de tussenliggende groep van genfamilies worden duplicaten tot wel meer dan tientallen miljoenen jaren na de volledige genoom duplicatie behouden, maar uiteindelijk keren deze genfamilies toch terug naar de één-kopie status. De oorzaak van dit patroon is mogelijk te wijten aan de dosis-sensitiviteit van deze genen. Dit mechanisme zorgt voor een langer behoud, maar kan na verloop van tijd opgeheven worden. Het onderscheid tussen de groepen van genfamilies wordt gereflecteerd in de functionele annotatie van deze groepen. De één-kopie genfamilies zijn voornamelijk verantwoordelijk voor het behoud van de genoom stabiliteit en de organel functies. De meerdere-kopij genfamilies zijn functioneel betrokken bij signalisatie, transport, en metabolisme en de tussenliggende groep bestaat voornamelijk uit regulatorische genen.

In hoofdstuk drie gaan we dieper in op de moleculaire mechanismen achter dit patroon. Hiervoor kijken we naar de impact van eiwit-eiwit interacties op het evolutionaire lot van volledige genoom en kleinschalige duplicaten in Arabidopsis, tomaat en mais. Door gebruik te maken van een duplicaat classificatie gebaseerd op fylogenetische bomen van genfamilies, grote RNAseq expressie compendia en een uitgebreid eiwit-eiwit interactie netwerk, konden we significante divergentie op sequentie, expressie en interactiepartner niveau waarnemen tussen kleinschalige en volledige genoom duplicaten. Daarnaast tonen we aan dat duplicaten die betrokken zijn bij eiwit-eiwit interacties sterker behouden zijn dan duplicaten zonder interacties in Arabidopsis, tomaat en mais. Duplicaten met eiwit-eiwit interacties zijn overgerepresenteerd in de

genfamilies die verondersteld worden dosis-sensitief te zijn en kunnen deels de patronen uit hoofdstuk twee helpen verklaren. Samenvattend toont dit hoofdstuk de invloed van eiwit-eiwit interacties aan op de evolutie van genduplicaten op basis van sequentie, expressie, interactie en genretentie data. Daarnaast tonen we aan dat zowel duplicatie type als interacties beschouwd kunnen worden als eigenschappen van genfamilies die het behoud van duplicaten beïnvloeden in alle bloemplanten.

TranSeq: 3'-einde sequenering

In hoofdstuk vier presenteren we TranSeq, een sequeneringsmethode die start vanaf het 3' einde van het mRNA. TranSeq vereist 10 tot 20 keer minder `reads` dan de huidige sequeneringsprocedures en verlaagt daardoor de kosten significant. Dit laat een grote stijging toe van het aantal stalen dat in één experiment geanalyseerd kan worden. In dit hoofdstuk bevestigen we dat deze methode accurate expressieprofielen genereert en verder gebruikt kan worden voor het verbeteren van de genoom annotatie. Het mappen van de TranSeq reads op het referentie genoom van tomaat heeft geleid tot een verbetering van meer dan 45% van de genmodellen. Samenvattend kan TranSeq de grootschalige transcriptoom studies stimuleren en de resolutie van genexpressie en genannotatie verbeteren in zowel model als niet model planten.

Geïntegreerde netwerk motief modules

In hoofdstuk vijf tonen we conceptuele inzichten aan in de topologie van eukaryote geïntegreerde genregulatorische netwerken dit respectievelijk in een functionele, dynamische en evolutionaire context. Verschillende types van moleculaire interacties werken nauw samen in deze netwerken om de correcte genen tot expressie te laten komen, op de juiste plaats en op het juiste moment. Omtrent hoe genen elkaar exact beïnvloeden zijn nog vele vragen onbeantwoord, zeker in hogere eukaryoten. Om inzicht hierin te krijgen hebben we een kader ontwikkeld om geïntegreerde genregulatorische netwerken te bouwen en te onderzoeken. Deze netwerken bestaan uit ongerichte (eiwit-eiwit, genetische, en homologe) en gerichte (eiwit-DNA, regulatorische, en miRNA-mRNA) interacties. Meer specifiek maken we gebruik van netwerk motieven en netwerk motief modules. Deze vormen topologisch biologisch relevante onderdelen van het genregulatorische netwerk in wormen en planten. Door integratie van expressie profielen werden dynamische netwerk modules bekomen. Via fylogenetische decompositie werden de netwerken ontleed wat toelaat de evolutionaire leeftijd van genen te linken aan hun topologische en functionele eigenschappen. Daarnaast bespreken we ook de invloed van specifieke data types en incomplete netwerken.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

BD: Birth death
BH: Benjamini Hochberg
CC: Clustering Coefficient
ChIP: Chromatin-Immuno Precipitation
CDS: Coding sequence
DPI: DNA-protein interaction
GF: Gene family
GMM: Gaussian mixture modelling
GO: Gene Ontology
GRN: Gene regulatory network
Kn: non-synonymous substitution rate
K-Pg: Cretaceous Paleogene
Ks: synonymous substitution rate
LOF: Loss of function
miRNA: microRNA
Mya: million years ago
PPI: protein-protein interaction
SCP: Single copy percentage
SSD: small-scale duplication
TAP: Tandem Affinity Purification
TE: transposable element
TF: transcription factor
WGD: whole-genome duplication
Y1H: Yeast-one-hybrid
Y2H: Yeast-two-hybrid

AIMS & THESIS OUTLINE

The duplication of genes and whole genomes is an important mechanism to increase evolutionary novelty. In plants, paleopolyploidy events (ancient WGD events) are found at the basis of all major lineages (e.g. angiosperms, eudicots and monocots) and within the domesticated crops polyploids have been selected for their higher yield and better fruits. Therefore, there is a need to study how genomes change after duplication events, how the resulting duplicates evolve over time, which molecular mechanisms influence duplicate loss and retention, and how these duplicates are integrated into the existing gene network context.

Chapter 1 introduces all necessary terminology and concepts regarding Omics data, gene regulatory networks and gene duplication in plants. This general basis will help understanding and interpreting the research chapters.

The first research chapter, Chapter 2, is a reproduction of a paper published in Plant Cell, in which we look at the different types of duplications and associated gene loss and retention patterns across angiosperms. Additional topics within this chapter are the various functional categories associated with the duplicate retention and loss, as well as the speed at which these losses occur and possible factors influencing either retention or loss. In general, this chapter is aimed at providing an overarching view of 'gene duplicability' across angiosperms.

The third chapter digs deeper into the influence of duplication mode (small-scale versus large-scale) and protein-protein interactions on loss and retention of duplicates by linking theories of duplicate retention like dosage balance sensitivity to observations from sequence, expression and interaction data of *Arabidopsis thaliana, Solanum lycopersicum* (tomato) *and Zea mays* (maize).

The fourth chapter presents a proof of concept using a novel 3' UTR sequencing technique, TranSeq. This sequencing approach aims to improve gene annotations and allows a cheaper and faster detection of gene expression.

The final research part (Chapter 5) proposes a novel pipeline to integrate multiple interaction data types and different experimental methodologies. The topological organization of integrated plant (*A. thaliana*) and worm (*C. elegans*) networks is studied by assembling the interactions through network motifs into network motif modules and adding a regulatory superview on top. Next to revealing topological features of biological networks this chapter is aimed at answering questions related to the evolution of networks over time and the integration of novel genes into the networks by using a combination of phylogenetic and structural decomposition. Overall, this chapter presents conceptual insights in the topology of eukaryotic integrated gene regulatory networks in a functional, dynamic and evolutionary context.

The last chapter (chapter 6) provides a general overview of the achieved goals, a discussion how this relates to published research, what questions are still open and what the next steps might be.

Enjoy the read!

1 INTRODUCTION

"The hand that gives is among the hand that takes. Genes have no fatherland, genomes are without patriotism and without decency, their sole object is gain."

Napoléon Bonaparte

1 INTRODUCTION

1.1 SETTING THE SCENE

The cells of all organisms on earth are built from four different types of molecules: nucleic acids, proteins, carbohydrates and lipids. According to the central dogma of molecular biology, the genetic code, encoded in the DNA (deoxyribonucleic acid), is transcribed into RNA (ribonucleic acid), which is then translated into proteins which carry out a broad range of functions. This "standard cellular pathway" is universal among eukaryote organisms. All molecules within a cell interact with each other and together they form a complex network, which is shaped by evolutionary pressure acting on the system. Through adaptation and evolution of this system, the overwhelming species diversity we see on our planet today was created. Species have adapted to occupy almost every niche on earth.

To learn more about biological systems and their evolution, specialized techniques were developed to detect, measure, and characterize each of these molecules. The continuous improvement of nucleotide sequencing and high-throughput detection techniques has made the detection and identification of most of these molecules affordable and accessible for the whole scientific community. This generates a still expanding information tsunami consisting of different types of -Omics data; each linked to a specific molecule or process within organisms (e.g. genomics, transcriptomics, metabolomics …). The most commonly used example of this tsunami is the exponential increase in the number of available genome sequences (DNA sequences in a species) (Figure 1.1). For plants there are 157 genomes available with varying assembly and annotation quality [1]. On top of that there are a magnitude of transcriptomes (all RNA sequences in a species) available. In recent years the genetic variation within a population is also being assessed by sequencing different accessions of the same species [2, 3] and other data types, for example, proteome (proteins) and interactome (interaction) data are also being gathered in high-throughput experiments.

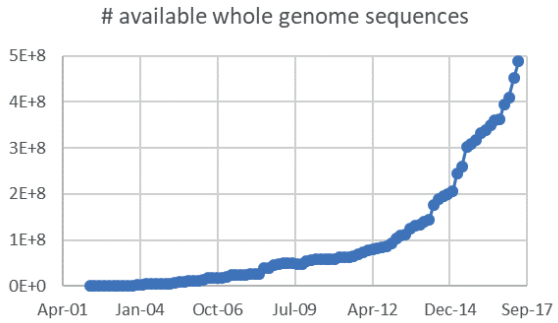


Figure 1.1: Increase in available genome sequences in the sequence read archive over time (source: SRA).

The huge amount of available data opens up a wide range of opportunities to study the molecular and evolutionary mechanisms which created the live we see today. This thesis, will take you through my journey studying the retention and loss patterns of gene duplicates in Angiosperm plants, the influence of protein-protein interactions on duplicate evolution, a novel 3'end RNA sequencing approach, and network motif modules in integrated gene regulatory networks.

1.2 EVOLUTION TROUGH MUTATION AND GENE DUPLICATION

The genome, which transfers the genetic information on to the next generations, undergoes frequent changes. On the nucleotide level mutations make changes in the DNA sequence. A first group of mutations are the substitutions in which of one base is exchanged for another one (E.g. A->G, C->T). If a substitution is located within a coding sequence (CDS) region it can be synonymous (silent mutation with no change to the amino acid it encodes), non-synonymous (mutation with a change to the amino acid it encodes) or non-sense (introducing an early stop codon). A second group of mutations are insertions and deletions, in which one or a few base pairs are inserted into or deleted from the DNA. This can cause a frameshift in the reading frame leading to wrong transcription and translation. For all mutations the effect is dependent on both the position and the type of the change [4]. Within the translated region they can have a direct effect on the protein function. If located outside the translated region the effect is dependent on the position of the mutation. Mutation within a cis-regulatory region for example can have a severe effect.

The biggest changes in gene content involve duplications, these can be small- or large scale. Duplicates drive evolutionary adaptations and novelty in genomes through introduction of new genetic material. This extra genomic content is essential for the creation of developmental functions and regulatory pathways [5]. In this part, you will find a general overview of the origin, the types, the use and the fate of duplicates. First, we have to introduce some terminology (Figure 1.2).

- **Homologs**: general term for two genes coming from a common ancestral DNA sequence, e.g. all genes on Figure 1.2 are homologs of each other.
- **Orthologs**: homologs separated by speciation event, e.g. Bα and Cα. A is an ortholog of all other genes.
- **Paralogs**: homologs separated by a duplication event, e.g. Cβ1 and Cβ2
- **Homeologs**: paralogs coming from a polyploidy event.
- **Gene family:** a set of genes that are thought to originate from the same gene and expanded through duplication and speciation. The genes can encompass one or multiple species and mostly have a similar function. E.g. all genes shown on Figure 1.2 form a gene family.

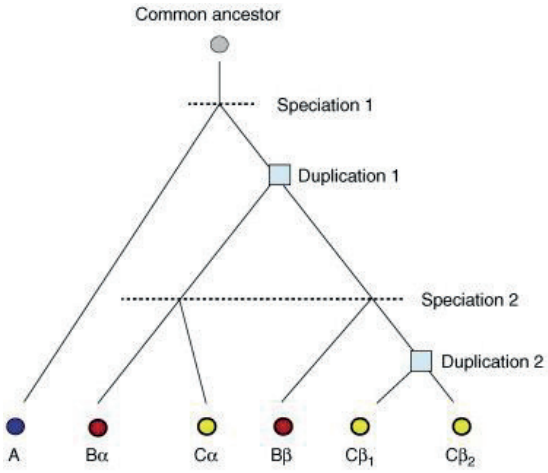


Figure 1.2: ***Gene family with marked duplication and speciation events.*** *The colours each represent a species.*

1.2.1 Small-scale duplications

There are different types of small-scale duplicates (SSD). The most prevalent ones are the tandem duplicates, which mostly originate from unequal crossing-over between two alleles or replication slippage (Figure 1.3A) [5]. Replication slippage is caused by mispairing of the DNA strands after denaturation [6]. This often produces small repetitive structures or partial gene duplications. Tandem duplication result in a cluster of paralogous sequences close to each other on the same chromosome. For example, the small heat shock protein gene family in tomato has expanded heavily through tandem duplication events [7]. SSDs can also be generated through transposon-mediated duplications (Figure 1.3B). In plants Pack-Mutator-like transposable elements (Pack-MULEs) induce the duplication of genes and genomic fragments into different genomic regions [8]. In mammals, another system with duplication through transposable elements called segmental duplication is active [9]. A third mechanism is retroduplication (Figure 1.3D). In this mechanism, the mRNA reversely transcribed to DNA and then re-inserted somewhere else in the genome [10]. The detection of SSD duplicates is commonly done with sequence similarity searches (e.g. BLAST search).

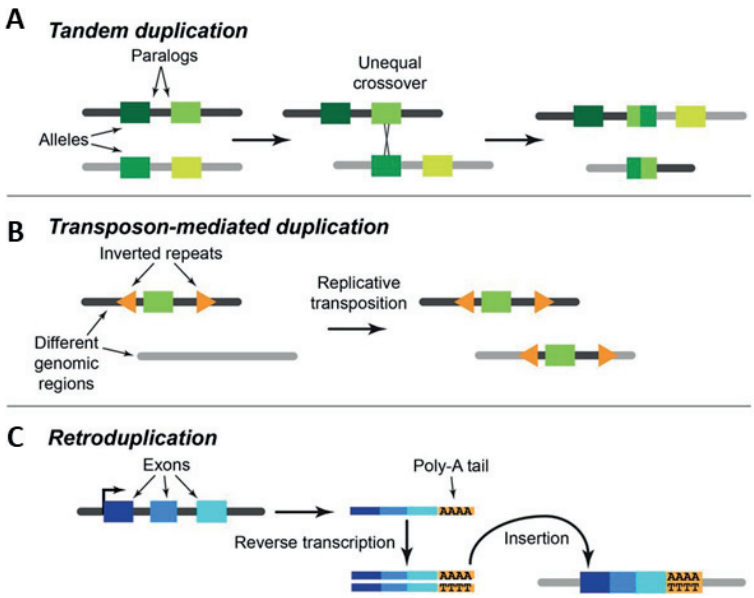


Figure 1.3: ***Overview of the SSD mechanisms.*** A) Tandem duplicates result from unequal cross-over of alleles. B) Transposon-mediated duplication of gene located within the transposon boundaries. C) Retroduplication. Reinsertion of a gene that underwent reverse transcription followed by insertion. This gene has lost intronic elements. *Figure from [5].*

1.2.2 Whole genome duplications

Polyploid species have multiple sets of chromosomes, this is the result of a WGD. Polyploids occur naturally but are also induced during crop-breeding. Polyploidy can either be an alloploidy or an autoploidy event. Alloploidy is the fusion of 2 or more chromosomal sets coming from a different species (interspecific hybridization), for example wheat is a cross between three grasses (see 1.2.3). In an autopolyploidy event, homologous chromosomal sets fuse together, this can be a doubling of the genome or a merge within the same species [11]. The merge of two genomes is claimed to give combinatorial benefits called hybrid vigour or heterosis [12, 13].

Despite the frequent occurrence of polyploids in natural populations, the establishment only happens rarely. It has even been argued that polyploidy is an evolutionary dead end [14, 15]. Still we detect many cases of paleopolyploidy (Figure 1.4) [16]. WGD have been found in the early vertebrate evolution (as Ohno predicted in the 70's [17, 18]), in fungi, amphibia and fishes, but WGDs are especially prevalent in the flowering plants (Figure 1.4) [16]. In model plant *Arabidopsis thaliana,* for example, at least four WGD events were detected. The Alpha (±50 Mya), the Beta (±60 Mya), the Gamma WGD which is shared between all Eudicots, and an older unnamed one shared between all angiosperms (Figure 1.4) [19-21]. The fact that so many ancient WGDs are detected means that at certain points in time the polyploids survived while the diploids became extinct. Many independent WGDs are detected around mass extinction events (Figure 1.4) [16, 22-24]. The link between WGDs and extinction events is however still debated, but it seems that stress can cause polyploidy and that under extreme environmental stress polyploids have an advantage [16, 22-25].

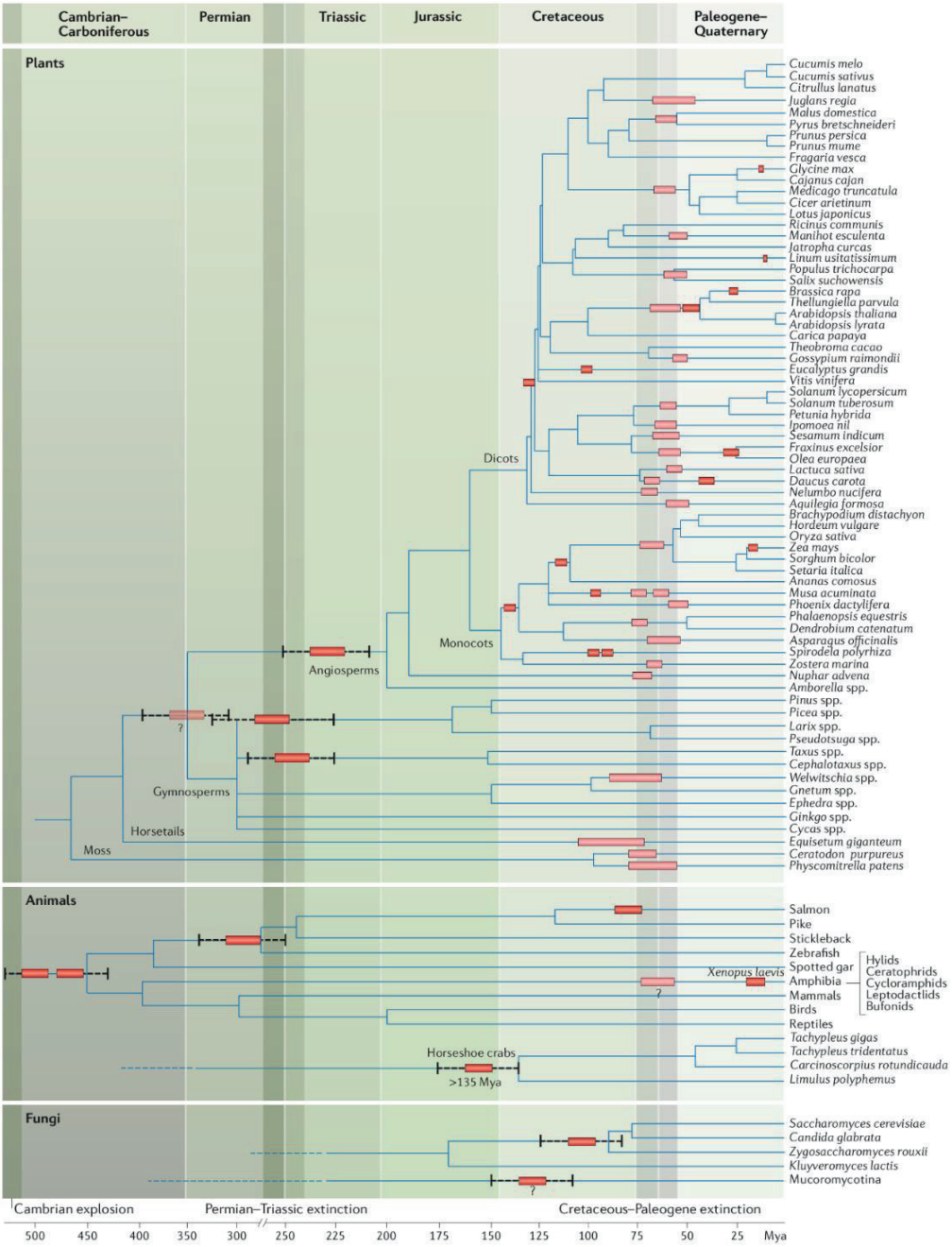


Figure 1.4: ***Phylogenetic tree with known whole-genome duplications.*** *The WGDs are marked in red on the tree with bold black dashed lines reflecting uncertainty in the date of the events. WGDs located around the Cretaceous–Paleogene boundary are marked in light red. The shaded areas represent mass extinction events. Figure from [16].*

Detection of WGDs

WGDs can be inferred using computational approaches either by detecting collinearity within a genome or by looking at the distribution of synonymous substitution rate (Ks) [26]. Often a combination of both is used.

Collinear regions are stretches of genes with conserved gene order and gene content between genomic segments. Homologous genes from these segments (multiplicons) are called block duplicates and they provide the evidence for large scale duplications (Figure 1.5A). Next to collinearity, Ks based age distribution between paralogs can be used to detect WGD events [27]. Synonymous substitutions (Ks) within a CDS region of a gene don't change the associated amino acid sequence, making them neutral due to the unchanged protein function. The fact that they appear at a constant rate (within a species) makes them suitable as a proxy for age. The age distribution for SSD is typically L-shaped with many recent duplicates and fewer older duplicates (Figure 1.5B), due to most duplicate genes getting lost. The other peaks showing a burst of gene duplicates, represent WGD events (Figure 1.5B). The underlying events can be found by fitting mixture models to the complete duplicate Ks based age distribution [27-29]. Combining the synonymous substitution rate with phylogenetic trees calibrated based on fossils makes it possible to estimate the date of the WGD event [21].

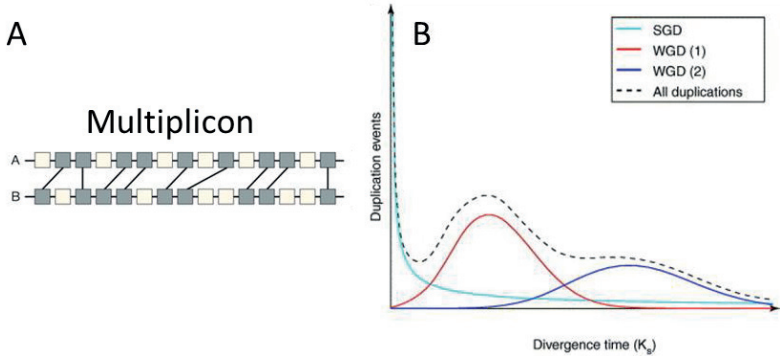


Figure 1.5 ***A: Multiplicon A and B represent two hypothetical genomic segments that are being evaluated for homology****. Each square represents a gene. Homologous genes are indicated in grey and connected by black lines in the collinear region (multiplicon). The homologous genes are also called anchor points and form the evidence for a segmental duplication. Figure adapted from [26].* ***B: Theoretical split up of the synonymous age distribution rate (Ks) into the different duplication events.*** *This example shows two whole-genome duplication events (red and blue) and ongoing small gene duplication events (SGD). If we add up all these distributions we get the complex full distribution [30].*

1.2.3 Importance of polyploidy in crop species

The expected population increase demands a higher food supply and in the same time the impact on the ecosystem needs to be reduced. To achieve this plant genes and genomes need to be characterized and genetic diversity needs to be explored [31]. Within angiosperms a wide range of paleoploidy events and a large amount of lineage-specific WGD events have been detected (Figure 1.4) [21]. These events are major force in the creation of adaptation and diversity and contributed to important agronomic traits [5]. Most of the crops underwent at least one round of WGD/polyploidy (Figure 1.6). Natural polyploids where domesticated and polyploidy was induced within crops. These plants were used within breading programmes that gave rise to the crops we grow and eat today. Figure 1.6 shows the importance of polyploids in evolution and domestication for wheat and Brassicaceae crops [31]. Multiple rounds of duplications gave rise to large complex genomes with a high number of homologs. Understanding the evolution of genome architecture, and how duplicated genes and genomes evolve over time, is important to improve agronomic traits of crops.

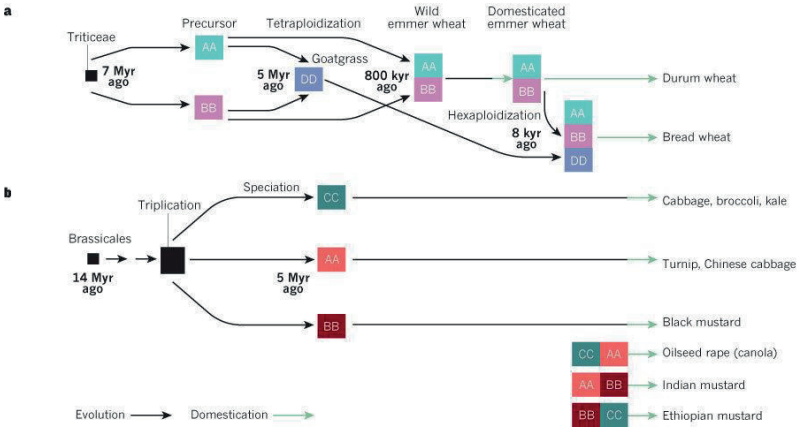


Figure 1.6: ***Influence of polyploidy in evolution and domestication of plants****. A) Polyploidisation and domestication history of wheat starting from the common ancestor, Triticeae, over the diploid precursors AA, BB and DD which gave rise to the tetraploid durum wheat and hexaploid bread wheat. B) Diverse Brassicaceae crops originating from the same ancestral Brassicales that underwent a genome triplication followed by speciation into 3 different species (AA, BB and CC). Those were domesticated and together with alloploid combinations they give rise to a wide range of crops that are grown worldwide. Figure from [31].*

1.2.4 Fate of duplicates

Within the angiosperms clade, WGD duplicates make up the largest amount of genome duplicates [20]. After the duplication event fractionation takes place, during which parts of the genome get lost. This results in a reduction of the genome size. Depending on the duplication mode, the fractionation can be biased towards one of the subgenomes [32]. The loss of duplicates is not restricted to fractionation. On the long term the most common fate of duplicates is loss (Figure 1.7) [5, 33, 34]. Loss can take place through increase of the mutational load which turns the gene into a pseudogene or the gradual reduction of the expression of one of the copies [35]. After this the gene can be removed from the genome without an effect on the fitness. On the other hand, retention of duplicates can be achieved through dosage of gene products, selection of the existing function, or creation of a novel function (Figure 1.7). The dosage related retention can either be due to dosage balance or absolute dosage of the duplicates. The dosage balance theory states that duplicates can be maintained because of stoichiometric dosage balance in complexes or pathways (relative dosage). Disturbing the balance between genes can have severe effects (e.g. malformed unfunctional protein complexes) [36-41]. The retention based on absolute dosage is related to the beneficial effect of a higher dosage of the gene product. This could be for example an enzyme being present in a higher concentration which leads to an increased flux through a pathway [42]. A second retention mechanism is sub-functionalisation where the original function is split between the two duplicate copies or the interfering function between the copies is resolved (paralog interference) [43]. The third mechanism is neofunctionalization, or the retention of duplicates because one copy gains an extra function which makes it beneficial for the organism. In this category we can also find the duplicates which escape from adaptive conflict in which both duplicates independently resolve their conflicting ancestral functions [5, 44].

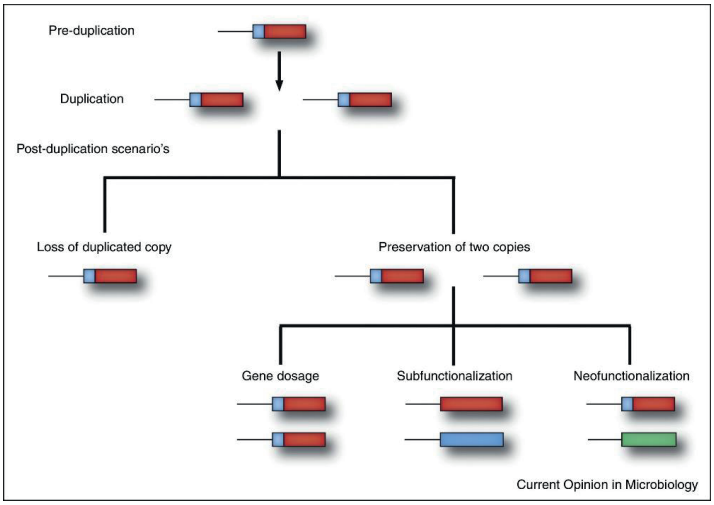


Figure 1.7: ***Potential fates of duplicate genes.*** *After duplication, there are different scenarios for the two copies of the duplicate. The most common one is loss of one of the copies. Duplicates are preserved because they have dosage related advantages or problems. They can split the function of the ancestral gene between the two copies (sub-functionalisation) or one of the copies can take on a new function (neofunctionalization). Figure form [45].*

It has been found that some duplicates are more easily lost while others are frequently retained. The retention of duplicates is biased in function and position (Figure 1.8) [46]. These biases can be observed in GO categories and duplication modes. Depending on the biological and environmental conditions certain functional categories can be maintained [47], similarly linked genes can be co-eliminated (e.g. [48]). Constraints related to dosage, regulation, expression or interactions can lead to duplication resistance [33]. Functional retention is linked to different duplication modes. After WGD, transcription factors, components of multi-protein complexes, and organellar genes are preferentially retained and after SSD involved in stress related pathways are retained [29, 49]. WGD duplicates are also asymmetrical lost [11] this has mostly been associated with allopolyploidy. Finally, reproductive isolation can lead to reciprocal loss of duplicates with loss of mating compatibility between the species [50, 51]. The retention forces which are active after duplication relax over time which leads to loss or change in fate of the duplicates. The paths are also not fixed paths, genes might jump from one into the other [48, 52].

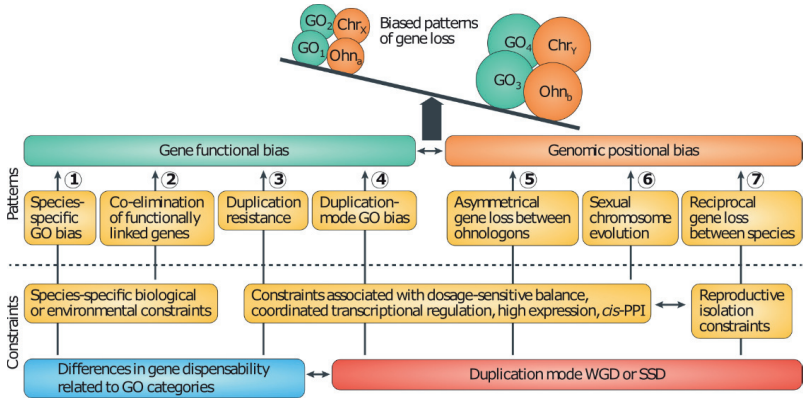


Figure 1.8: ***Patterns and constraints leading to prolonged gene retention and loss after duplication****. Genes can be lost or kept due to functional (green) of positional (orange) biases which can originate from constraints related to gene ontology (blue) or duplication mode(red). Biological and environmental constraints can lead to retention of certain GO categories (1) or co-elimination of linked genes (2). Constraints related to dosage, regulation, expression or interactions can lead to duplication resistance (3), link between duplication mode and function (4), asymmetrical loss of WGD duplicates (5) or sex chromosome evolution (6). Reproductive isolation can lead to reciprocal loss of duplicates. Figure from [46].*

1.3 THE INTEGRATED GENE REGULATORY NETWORK

Within and surrounding the cells of organisms there are huge number of functional molecules. All these components work together to organise the whole metabolism of the organism, keep it stable, and allow it to respond to internal and external factors. All the molecules (nodes) and the interactions between them (edges) form multi-layered complex networks. The most common types of biological networks which involve genes and proteins are: protein-protein interaction networks (undirected network representing the physical relationships between proteins), genetic interaction networks (undirected network showing functional relationship between genes) and gene regulatory networks (directed network representing gene regulation). Within this thesis we group these three networks involving genes and proteins together into the integrated gene regulatory network (GRN) (Figure 1.9) [53]. All these components interact with each other and regulate directly or indirectly how the genomic DNA content produces the correct amount of RNA, proteins and metabolites under specific conditions. Other biological networks which are less relevant for this thesis are metabolic networks (biochemical reactions between enzymes and metabolites) and Cell signalling networks (pathways combining gen regulatory and metabolic networks).

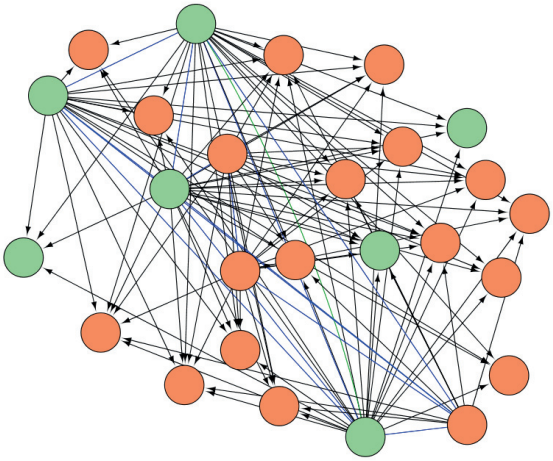


Figure 1.9: ***Example of an integrated gene regulatory network.*** *Green nodes are regulators (e.g. TFs) and orange nodes are target genes. This example contains directed regulatory edges (black; e.g. protein-DNA interactions) and undirected edges between genes (blue and green; e.g. protein-protein interactions or genetic interactions).*

1.3.1 Regulators

Within the GRN there are many regulators, each with their own characteristics and functions. In this thesis we will focus on TFs and miRNAs (Figure 1.9). Other regulation types which will not be discussed here are for example epigenetic regulation [54], regulation through RNA binding proteins [55], and post-translational protein modification [56].

TFs are DNA binding proteins that activate or inhibit transcription of genes by binding to the DNA in the promotor region or further away at enhancer regions of a gene. They can work in the form of dimers, together with co-factors or chromatin modifying proteins to regulate very specific or broad cellular processes [57]. Dimers are a combination of two TFs, if these are identical they are called homodimers and in the case of different TFs they are called heterodimers.

MiRNAs are small RNA molecules (20-24 nucleotides) with a very specific structure. They silence the expression by binding to the target site of the mRNA. The mature miRNA is capsuled in the RNA-induced silencing complex (RISC). MiRNAs belong to a large group of RNA derived molecules which also contain, small interfering RNAs (siRNAs), Piwi-associated RNAs (piRNAs) and long non-coding RNAs (ncRNAs) [58, 59]. They vary a lot in size, structure, and mechanism of action; but all of them are proven to be essential parts of the gene regulatory network.

1.3.2 Molecular interactions: types and detection

Biological networks consist out of a wide range of interactions which can be detected using high- or low-throughput experimental techniques, computational approaches [60], or extracted from literature using text mining algorithms [61]. For most interaction types, there are specific databases which gather all the publicly available data (e.g. STRING for protein-protein interactions [62]). In the last decade the amount of available interactions has increased largely due to specific efforts of consortia (e.g. Arabidopsis interactome [63]), and big projects to characterize all the elements and interactions present in the genomes of model organisms

(e.g.: ENCODE for human [64] and modENCODE for *Drosophila melanogaster* and *Caenorhabditis elegans* [65]).

Each interaction type has its own characteristics and detection methodology. Protein-protein interactions (PPI) can form multi-subunit complexes or transient interactions for regulatory functions such as signalling, or during the modification, degradation, or folding of other proteins. The most used techniques for the detection and identification of PPIs are Yeast-two-hybrid (Y2H) and Tandem Affinity Purification (TAP) (more info [60, 66, 67]). Proteins can also interact with DNA in protein-DNA interactions (PDI). In this category we have TFs, chromatin modifying proteins, polymerases and nucleases. Most of these proteins have specific DNA interaction domains (e.g. zinc finger). PDIs can be experimentally detected using a transcription factor (TF) centred approaches such as chromatin-immuno precipitation (ChIP) in which one TF is crosslinked to all binding sequences, or with gene centred approaches such as yeast-one hybrid (Y1H) in which all binding partners of a specific genetic region are detected. Prediction of PDIs can be done by searching known binding motifs in the promotors of genes or by looking at the change in expression in knock-out mutants. More information about the techniques and prediction methods can be found in [68, 69]. Not all interactions involve proteins directly, an example of this is miRNA-mRNA interactions. Numerous miRNA-mRNA interactions have been experimentally verified and wide range of prediction tools for these have been developed [70, 71]. A Genetic interaction originate if the effect of knocking out two genes is unexpected compared to the individual gene knock outs, which shows that there is a functional relationship between the genes and/or the pathways [72]. This direct or indirect link can come from synergistic effects. An example of this is synthetic lethality where the double knock-out is lethal while none of the individuals independently is lethal [73]. Next to these, other interaction types such as protein-RNA, RNA-RNA and epigenetic interactions are present within a cell. There might still even be unknown molecular interactions inside the complex cellular environment.

1.3.3 The biological network structure and applications

Networks are usually described using a specific terminology. The degree is the number of edges connected to a node. Nodes which have a high degree are called hubs. The Betweenness of a node is the number of shortest paths between two nodes that pass through that node. Biological networks have a specific structure. They tend to be 'scale-free', 'small-world' and show a high modularity [74]. In scale-free networks, the degree follows a power-law distribution. This means that there many of nodes with only a few interaction partners and few nodes with many of interaction partners. In small-world networks the average path length between two nodes is short. High modularity points to the fact that the complete network consists out of many smaller network clusters and is expressed in the clustering coefficient (CC). Nodes with a high degree, betweenness and clustering coefficient are found to be more central in the network and are thought to be more essential [75].

Within the network there are smaller structures like network motifs or modules [74]. Network motifs are small, well defined, network patterns which appear frequently (Figure 1.10). They usually consist out of two to five nodes and can be seen as the small building blocks of the network. The detection happens through enrichment analysis in real versus random networks. The best known and most abundant motif is the feed-forward loop, where a regulator regulates a target gene directly and indirectly through another regulator (Figure 1.10; FFL). Other types of motifs have been described in multiple organisms (Figure 1.10) [76-81].

Where motifs have a small number of nodes with a defined structure, network modules are sets of genes without a defined structure of size. They can be up to a couple of hundreds of genes. These genes can be assigned into modules based on network or gene characteristics such as interaction density, function or expression similarity.

Biological networks, motifs, and modules can be used for many different applications [82]. A first one being the prediction of the function of a certain gene by looking at its context. Vice versa it is also possible to determine new genes functioning in a certain process (e.g. [83]). Through networks it is also possible to explain or predict the outcome of gene perturbations or predict the relationship between genes. The latter can be useful for predictions related to disease or mutant conditions (e.g. [84]).

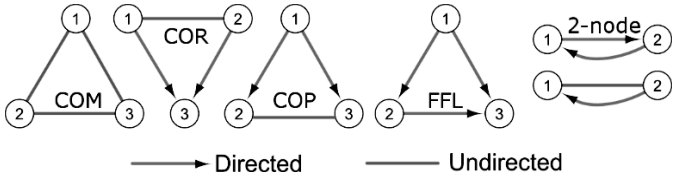


Figure 1.10: ***Network motifs.*** *Examples of 2 and 3-node network motifs composed out of directed and undirected edges. COM: three nodes connected through undirected interactions (e.g. protein-protein interactions). COR: two interacting regulators that regulate the same gene (e.g. dimers). COP: co-pointing motif, a regulator regulating two connected genes. FFL: feed-forward loop, where a regulator regulates a target gene directly and indirectly through another regulator.*

1.3.4 Measuring the RNA level

The activity of genes in GRNs is mostly measured by quantifying the RNA expression level, which can be done using microarrays or RNA sequencing (RNAseq). Microarrays are composed of probes attached to a solid plate (mostly glass). These probes are a set of short nucleotide sequences which are representative for a set of genes [85, 86]. For this, the sequence of the species under investigation has to be known. The RNA abundance of the probe set is determined by hybridising fluorescently labelled transcripts and measuring the intensity of each probe [87]. Since a couple of years RNAseq has overtaken microarrays as the most used transcriptional technique [88]. This technique, based on high-throughput sequencing, allows for the deep sampling of the whole transcriptome and is not limited to a set of predefined genes as is the case with microarrays. The generated reads can be aligned to a known genome/transcriptome or can be de-novo assembled [89]. This allows to quantify genes under specific conditions and enables the identification of novel genes. The ever-proceeding challenge for higher specificity and accuracy has now enabled to the detection of the transcriptome of single cells [90]. Next to measuring RNA levels it is also possible to quantify the protein expression levels using proteomics [91, 92].

1.3.5 Functional annotation of genes using gene ontology

Gene ontology (GO) is an initiative to make a uniform vocabulary to functionally annotate genes and their products [93]. It is developed as a species neutral machine-readable data format to enable easy access and use for functional interpretation of experimental data. The terms, called ‘ontologies’, are split into three domains: cellular component (part in or out the cell), molecular function (activity at molecular level) and biological process (operations or sets of molecular events). GO is structured as a directed acyclic graph with defined relationships between them. The most known usage of GO is enrichment analysis of gene sets [94].

Through orthology the functional knowledge can be transferred to other species. This is for example done for plants in the comparative genomics platform PLAZA [1].

2

GENE DUPLICABILITY OF CORE GENES IS HIGHLY CONSISTENT ACROSS ALL ANGIOSPERMS

"Loss is nothing else but change, and change is Nature's delight" *- Marcus Aurelius*

2 GENE DUPLICABILITY OF CORE GENES IS HIGHLY CONSISTENT ACROSS ALL ANGIOSPERMS

This chapter was published in The Plant Cell, Vol. 28: 326–344, February 2016

Zhen Li*, **Jonas Defoort***, Setareh Tasdighian, Steven Maere, Yves Van de Peer, Riet De Smet

*Equal contribution

2.1 ABSTRACT

Gene duplication is an important mechanism for adding to genomic novelty. Hence, which genes undergo duplication and are preserved following duplication is an important question. It has been observed that gene duplicability, or the ability of genes to be retained following duplication, is a non-random process, with certain genes being more amenable to survive duplication events than others. Primarily, gene essentiality and the type of duplication (small-scale versus large-scale) have been shown in different species to influence the (long-term) survival of novel genes. However, an overarching view of 'gene duplicability' is lacking, mainly due to the fact that previous studies usually focused on individual species and did not account for the influence of genomic context and the time of duplication. Here, we present a large-scale study in which we investigated duplicate retention for 9,178 gene families shared between 37 flowering plant species, referred to as angiosperm core gene families. For most gene families, we observe a strikingly consistent pattern of gene duplicability across species, with gene families being either primarily single-copy or multi-copy in all species. An intermediate class contains gene families that are often retained in duplicate for periods extending to tens of millions of years after whole-genome duplication, but ultimately appear to be largely restored to singleton status, suggesting that these genes may be dosage balance-sensitive. The distinction between single-copy and multi-copy gene families is reflected in their functional annotation, with single-copy genes being mainly involved in the maintenance of genome stability and organelle function and multi-copy genes in signalling, transport and metabolism. The intermediate class was overrepresented in regulatory genes, further suggesting that these represent putative dosage-balance sensitive genes.

2.2 CONTRIBUTION

- Performing the research together with Zhen Li and Setareh Tasdighian
- Designing and performing analyses on gene family data, gene family evolution and gene function
- Figures: 2.3, 2.4, 2.5, 2.6, S2.6, S2.7, S2.8, S2.9, S2.11 and S2.12
- Tables: S2.1, S2.2 and S2.3
- Assisting in writing the manuscript

2.3 INTRODUCTION

Since the seminal work of Susumu Ohno [17], the importance of gene and genome duplication for evolution and adaptation has been well-appreciated. Indeed, ample examples of gene diversification following duplication have been described and 'gene duplicability', by which we mean the ability of genes to be preserved in a population following duplication, has been extensively studied [95-101]. Studies published on a large array of species seem to converge on the idea that some duplicated genes are more likely to be preserved in a population, and as such to potentially contribute to functional innovation, than other genes. One factor that seems to influence gene duplicability is the mode of duplication, as in several organisms that have undergone ancient WGD it has been shown that different sets of genes were retained following WGD and SSD events [29, 102-106].

Both SSD and WGD have occurred frequently in the flowering plant lineage, and in particular WGDs have happened at a much higher rate than in, for instance, fungi or animals [107, 108]. Studying the Arabidopsis genome, it has been observed that certain sets of genes have almost exclusively duplicated through WGDs [29, 105, 106]. These genes have distinctive functional features, as they primarily encode TFs and components of multi-protein complexes, and are involved in development and in signalling pathways [29, 104-106]. A potential explanation for this phenomenon is given by the 'gene dosage balance theory', which states that for many genes that participate in essential complex cellular networks or protein complexes, it is crucial that the stoichiometry between the gene products is maintained [102, 103, 109-112]. While WGD preserves the relative dosage between genes, the stoichiometry is disrupted when only one or few interaction partners are duplicated. In other plant species, vertebrate and unicellular organisms that have also undergone ancient WGDs, similar observations were made [102, 113-116]. Hence, while gene loss following SSD is generally a relatively fast process, with average duplicate half-life estimates being in the range of a few million years [28], after WGD, a substantial set of genes is often retained in duplicate for a much longer time [29]. For instance, it is estimated that about 16% of the genes for *A. thaliana* are still present in duplicate following the most recent WGD that occurred about 49 Mya [108], while 75% of the genes are still present in duplicate in soybean, which underwent a WGD approximately 13 Mya [117]. Whether these genes will be retained indefinitely is still an unresolved question [118-120], although the lower numbers of retained genes reported for more ancient WGD events seems to suggest that, at least for a subset of genes, dosage constraints eventually get relaxed, leading to functional diversification or loss of these genes.

In stark contrast to observations of prolonged retention of a set of 'dosage-sensitive' genes are recent observations that a substantial fraction of 'core angiosperm genes', i.e. genes that are present in all angiosperm genomes, occur as singletons throughout, suggesting that their duplication might be detrimental [33, 112, 121-124]. While these observations are not necessarily in contradiction with each other, as they likely concern different gene sets, an overarching picture that unifies the different observations regarding 'gene duplicability' is currently still missing. Specifically, the fact that most studies concerning 'gene duplicability' report species-specific patterns adds to the confusion, as genetic context, species biology, ecological requirements at the time of duplication and the timing of the WGD event might greatly influence the observed duplicate retention patterns [125-128].

Here we undertake a large-scale comparative approach to determine whether patterns of gene duplicability can be generalized across diverse lineages. In particular, we investigate the duplicability of 9,178 core angiosperm genes identified across 37 different angiosperm genomes and covering 20 putative WGD events. For most gene families, our analyses reveal a striking non-random picture of gene duplicability, with the majority of the core genes occurring as single copies in almost all of the angiosperm genomes and a more restricted set of genes occurring in duplicate throughout. This pattern is supported by a strong functional dichotomy between both classes of gene families, with single-copy genes being involved in the maintenance of genome integrity and organelle function, and multi-copy genes being biased towards signalling, transport and metabolism. Next to these two extremes, we also identified an intermediate class of gene families that show a pattern of prolonged duplicate retention spanning several tens of millions of years following WGD but appear to eventually also mostly return to singleton status. We hypothesize that dosage-balance constraints prolong duplicate retention in these particular gene families. Overall, we advocate that, at least for genes present in all angiosperms, the so-called core genes, selection plays an important role in the long-term preservation or non-preservation of duplicated genes, considering the highly non-random pattern that arises in this cross-species and cross-duplication event analysis.

2.4 RESULTS

2.4.1 Core angiosperm gene families show a strong preference towards the single-copy state

We collected the protein coding sequences for 37 sequenced angiosperm genomes (Figure 2.1) and constructed gene families using OrthoMCL (see Materials and Methods). To ensure that each of these gene families traced back to a single angiosperm ancestral gene we further processed these gene families using phylogenetic tree construction followed by reconciliation of the gene trees and the species tree (see Materials and Methods). Of the 69,133 gene families that were obtained using OrthoMCL and verified by phylogenetic analysis, 9,178 belong to the angiosperm core genome, defined as that part of the genome containing genes present in all angiosperms, including the angiosperm ancestor. To accommodate for errors in genome annotation, the presence of partial genome sequences and errors in gene family construction and/or phylogenetic analysis, we allowed for gene families in this core set to be missing in up to five genomes (see Figure S2.1 for a justification of this threshold). This set of genes was used in this study for all subsequent analyses. For each gene family, we calculated the fraction of species for which the gene family contains exactly one copy, further referred to as 'Single-Copy Percentage' (SCP). For instance, a value of 0.7 means that for that particular gene family, 70% of the species examined have exactly one copy while 30% of the species have more than one copy. The distribution of the SCPs for all core gene families is depicted in Figure 2.2. As can be observed, the distribution is highly skewed towards high SCPs, with the mean of the distribution lying at 66.8% and the mode of the distribution at 87.5%. Furthermore, if we remove genomes that still have a high number of retained duplicates due to a recent (< 20 mya) WGD event (such as soybean,

flax, maize and *Brassica rapa*, Figure 2.1), we observe an even stronger shift towards the single-copy state with the mode of the distribution being at 92.5% (Figure S2.2).

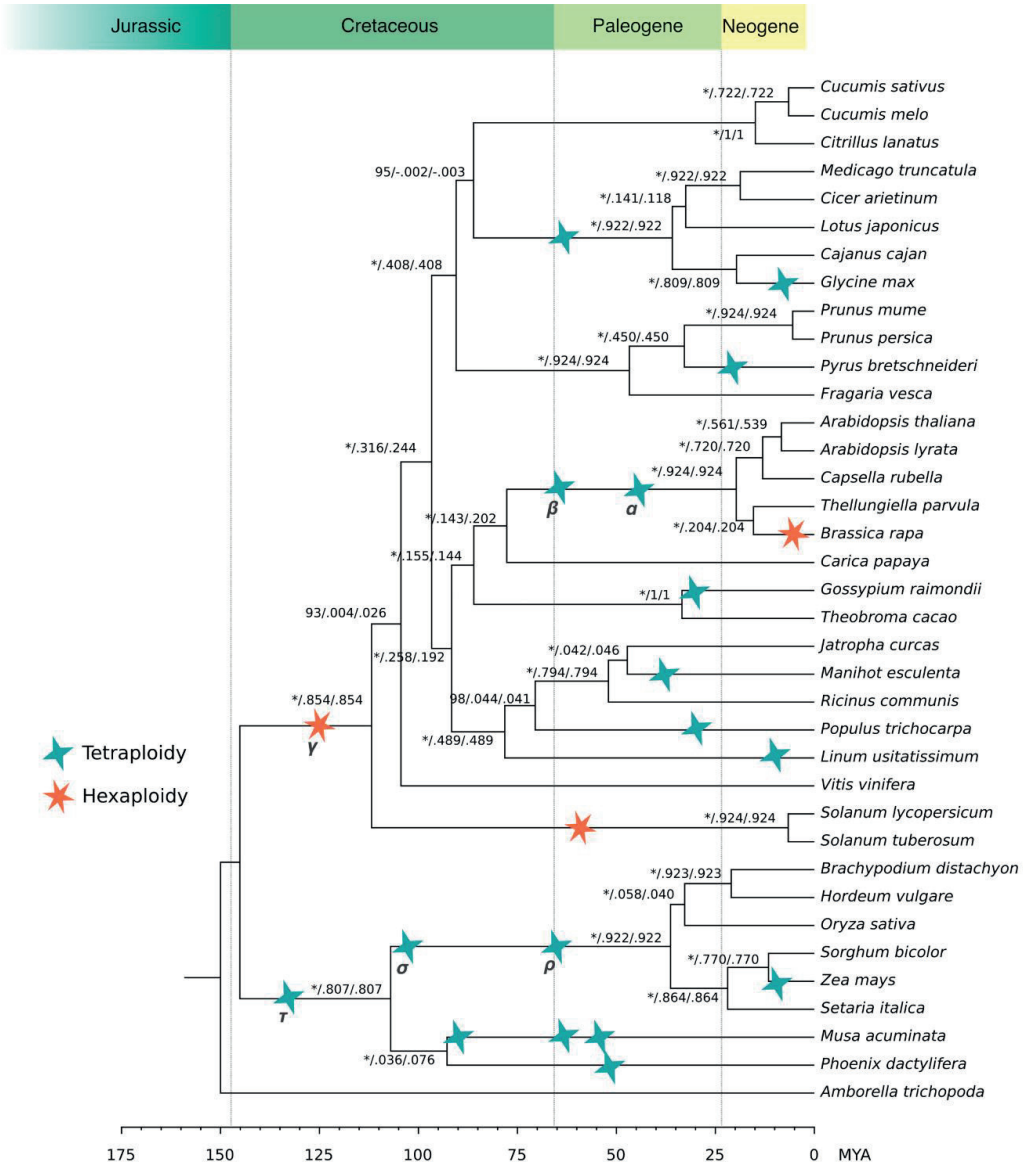


Figure 2.1: ***Angiosperm species tree.*** *Phylogenetic tree depicting the relationships amongst the 37 angiosperm genomes used in this paper. The tree topology was inferred from a concatenated alignment based on 107 almost single-copy gene families (see Materials and Methods). Numbers on the branches represent bootstrap supports, internode certainty (IC) and internode certainty all (ICA), respectively. Whole-genome duplication (WGD) events were inferred from literature [108, 129] and are depicted by stars. Only WGD duplications were considered that are more recent than the angiosperm common ancestor.*

Since the most likely outcome following gene duplication is duplicate loss, with average duplicate half-lives estimated at a few million years for SSDs [28], we have assessed whether our observations could be explained by simple stochastic gene duplication and loss dynamics. Therefore, we simulated gene family copy-number

evolution along the 37 species tree, using a probabilistic model in which SSD is modelled as a random birth-death (BD) process [130] and that takes into account known WGD events by assuming an instantaneous doubling (or triplication) of all genes, as in Rabier, Ta [131] (see Materials and Methods). Using this model as a null hypothesis and using realistic rates of SSD and loss, λ, sampled from a normal distribution with mean μ = 0.53 and standard deviation σ = 0.156 duplications/losses per evolutionary time unit (see Materials and Methods), we generated gene counts at the leaves of the species tree for 9,178 x 1,000 = 9,178,000 simulated gene families. We observe that the SCP distribution under the null model has a mode of 22.5% on average, compared to 87.5% for the core angiosperm gene families and that both distributions are significantly different ($p < 2.2e-16$, Wilcoxon rank-sum test) (Figure 2.2). Hence, under the neutral scenario of stochastic gene birth and death, there is no bias towards the single-copy state. We have repeated this analysis for different sampling distributions of λ-values and observed that the general trend of the distribution of SCPs for the simulated families remains similar, indicating that rejection of the null hypothesis is robust with respect to changes in the distribution of λ-values. Therefore, our observations suggest that gene families belonging to the so-called angiosperm core genome (i.e. gene families present in all angiosperm genomes) are skewed towards the single-copy state more strongly than expected under a random gene duplication-loss process and hence appear to be under (strong) selection to be single-copy.

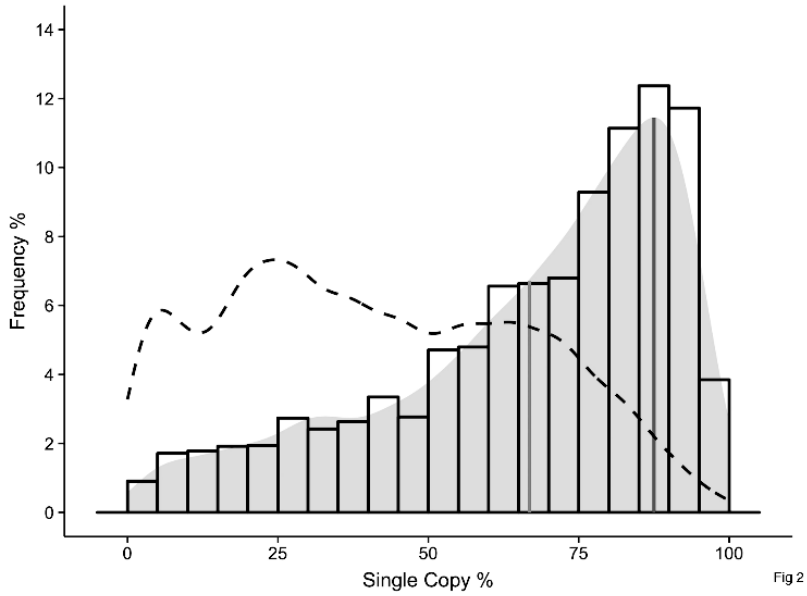


Figure 2.2 ***Overall distribution of single-copy percentage for all angiosperm core gene families.*** *The distribution depicts the degree to which the 9,178 core gene families are single-copy in the 37 angiosperm species investigated. The x-axis represents, for each gene family, the percentage of species with exactly one gene copy with respect to the total number of species in the family. The distribution illustrates a very strong tendency of angiosperm core gene families towards the single-copy state. The mode (87.5%) and the mean (66.8%) of the distribution are indicated by green and red lines, respectively. The observed distribution strongly deviates from the expected distribution under a stochastic duplicate birth-death model (depicted by dashed lines).*

2.4.2 Homeologs are quickly lost following WGD

The observation that many core gene families are single-copy, in spite of the large number of both recent and ancient genome duplication events, seems to suggest that gene loss occurs relatively fast following WGD. The large number of WGD events in this study and their different ages (Figure 2.1) provide an excellent case to study duplicate retention following WGD [132].

To study the dynamics of duplicate gene retention in the core gene families, we first assessed the contribution of WGDs as compared to SSDs to duplicate retention in the core gene families. Specifically, we applied gene tree - species tree reconciliation to obtain predictions of duplication events and their associated timing for all gene families (see Materials and Methods). To this end, we classified each node in the species tree (Figure 2.1) as either being associated with WGD or SSD, based on whether WGD events have been predicted on the branch leading to the specific node (Figure S2.3). Then we compared the predicted numbers of duplication events at WGD nodes versus SSD nodes for both core and non-core gene families, the latter referring to gene families that arose more recently than the angiosperm common ancestor or that underwent massive gene loss in some species since speciation from the angiosperm common ancestor. For the core gene families, we estimated that in total 69.8% (65,531 out of 93,942 predicted duplication events) of the duplications could be attributed to WGDs, whereas for the non-core gene families this was only 34.6% (48,778 out of 140,786 predicted duplication events) (Figure S2.4). Hence, for core families, as compared to non-core gene families, the presence of duplicates seems to be biased towards WGD-associated gene duplication ($p < 2.2e-16$, Fisher's exact test). In further support of the hypothesis that core gene families were more heavily impacted by WGD than non-core gene families, we observed that KS (number of synonymous substitutions per synonymous site)-based age distributions of duplicated gene pairs in the different species show clear peaks for the predicted WGD events if only duplicates from the core gene families are considered, while these peaks seemed to be absent for age distributions constructed for duplicates of non-core gene families (Figure S2.5). Hence, core gene families appear to be particularly suited to study duplicate preservation patterns following WGD.

We took advantage of the large number of WGD events and their different ages to study the dynamics of gene duplicate loss following WGDs. To this end, we assigned retained duplicates in the core gene families to the different WGD events or as being created by SSD based on a Gaussian Mixture Modelling (GMM) approach (see Materials and Methods). This way, for each species we obtained predictions of the timing (expressed in KS-values) of the WGD events they experienced and the number of gene families with retained duplicates for each of the WGD events [133-135] (see Materials and Methods). We used these data to assess the relationship between the number of gene families with retained duplicates and the estimated timing of the WGD events. As can be seen in Figure 2.3, duplicate retention subsequent to WGD follows an L-shaped curve that can be approximated by a power-law function (see Materials and Methods), confirming common expectations that gene loss subsequent to WGD is initially fast and then slows down. A similar power-law pattern was recently also observed in a genome-wide analysis of duplicate retention following WGD for a more restricted set of genomes [132]. For ease of interpretation, we grouped the WGD events into three different sets according to the overall time frame during which the WGD event occurred. ‘Ancient’ refers to the WGD events that have been predicted to have occurred at least 75 million years ago (Figure 2.1). This includes the ancient WGD event that is shared by all dicots and the WGD event that is shared by the Poaceae.

Using the mixture modelling approach, we could not find support for the predicted ancient WGD event that is shared by all monocots [129]. 'K-Pg boundary' refers to WGD events situated at approximately the K-Pg (Cretaceous-Paleogene) boundary, which reflects a clustering of WGD events at approximately 50-70 mya [108]. Finally, the 'recent WGD' set includes the duplication events that are more recent than the K-Pg boundary (< 50 mya). In Figure 2.3, duplicate retention patterns associated with the 'recent WGD' events show a steep decline as a function of WGD age. Whereas on average 41.64% (s.d. 21.74%) of the core gene families retain duplicates for the recent WGD events, for the 'K-Pg boundary' WGDs the number of core gene families with retained duplicates has dropped to on average 16.04% (s.d. 7.48%), and for the 'Ancient set' this number further reduces to 8.37% on average (s.d. 2.24%).

The distinction between SSD and WGD duplicates in this paper are approximate and SSD numbers are likely underestimated by both strategies (GMM and reconciliation method), because some SSDs might be located on a WGD branch (gene tree – species tree reconciliation) or might be hidden under a WGD peak (GMM analysis). However, we do not expect this to have a large influence on the observations that core gene families in contrast to non-core gene families are mainly duplicated by WGD nor on observed differences in gene duplicability patterns for different gene family groups (see further), as this underestimation likely affects all gene families equally.

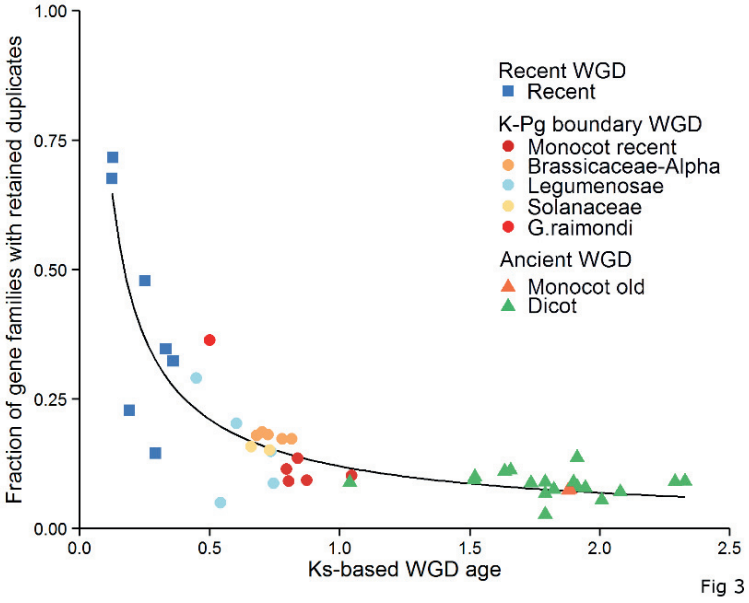


Figure 2.3: ***Duplicate gene retention in function of time since WGD.*** *Each dot represents the fraction of core gene families with retained duplicates following a specific WGD (y-axis), as a function of WGD age, expressed in KS-units (x-axis). The timing of the WGD events and the particular gene families that retained duplicates following a specific WGD event were inferred by fitting Gaussian mixture models to KS-age distributions for all 37 species separately (see Materials and Methods). As such, each point represents a species-specific estimate for a WGD and WGD events shared by multiple descendant species will be represented by multiple data points that cannot be regarded as being independent. SSD-related peaks and dubious WGD peak callings were omitted. Additional information on all the peaks can be found in Table S1 and Figure S7. A power-law function was fitted to the data (Chi-squared goodness-of-fit = 0.77, p = 1).*

2.4.3 Core gene families belong to different groups that reflect major differences in gene duplicability

Our global analyses on duplicate retention following WGD show that the majority of the angiosperm core gene families revert quickly to the single-copy state following WGD. Yet, the distribution in Figure 2.2 suggests that certain gene families revert faster to single-copy status than others. Therefore, we explored gene family specific differences in duplicate retention by constructing a copy-number profile matrix, which for each gene family lists the number of genes for a given species. We classified gene families into different groups based on an unbiased clustering of their copy-number profiles. By using a sub-sampling strategy in combination with clustering [136] (see Materials and Methods) we found that the data are best described by three stable clusters (Figure 2.4A, Figure S2.6, Figure S2.7): Group 1 contains 5,097 gene families and covers 5,473 *A. thaliana* genes, Group 2 contains 2,832 gene families and covers 4,312 *A. thaliana* genes and Group 3 contains 1,249 gene families and covers 3,255 *A. thaliana* genes. The heatmap in Figure 2.4A clearly shows the overall tendency of gene families in Group 1 to occur as single copies. If duplicates are present these are mainly biased towards species with recent WGDs. Gene families within Group 2 show mainly duplicate retention for species that are associated with 'Recent' and 'K-Pg Boundary' WGDs, while being largely single-copy for species that only underwent 'Ancient' WGDs. The latter suggests that while duplicates for these gene families are in general preserved for prolonged times, they eventually largely return to single-copy status. Finally, gene families in Group 3 have retained duplicates for all species, also for the ones that only underwent 'Ancient' WGDs. We also observe that the outgroup species Amborella trichopoda, which has no evidence of WGDs postdating angiosperm diversification [137], seems to be singleton for most of the core gene families, further substantiating the above observations that core gene families mainly duplicate through WGDs. Investigating the SCPs for the gene families in the three groups confirms that gene families in the first group show a strong preference towards the single-copy state, whereas gene families in the third group represent gene families with a strong tendency to be multi-copy in the majority of the species. The SCP distributions for each of the three groups are significantly different ($p < 2.2e\text{-}16$ for all comparisons, Kruskal-Wallis test followed by Dunn's test with Benjamini-Hochberg multiple testing correction) and there is almost no overlap in SCPs for Group 1 and Group 3 (Figure 2.4B). We will further refer to the gene families in Group 1 as 'Single-copy', those in Group 2 as 'Intermediate' and those in Group 3 as 'Multi-copy'.

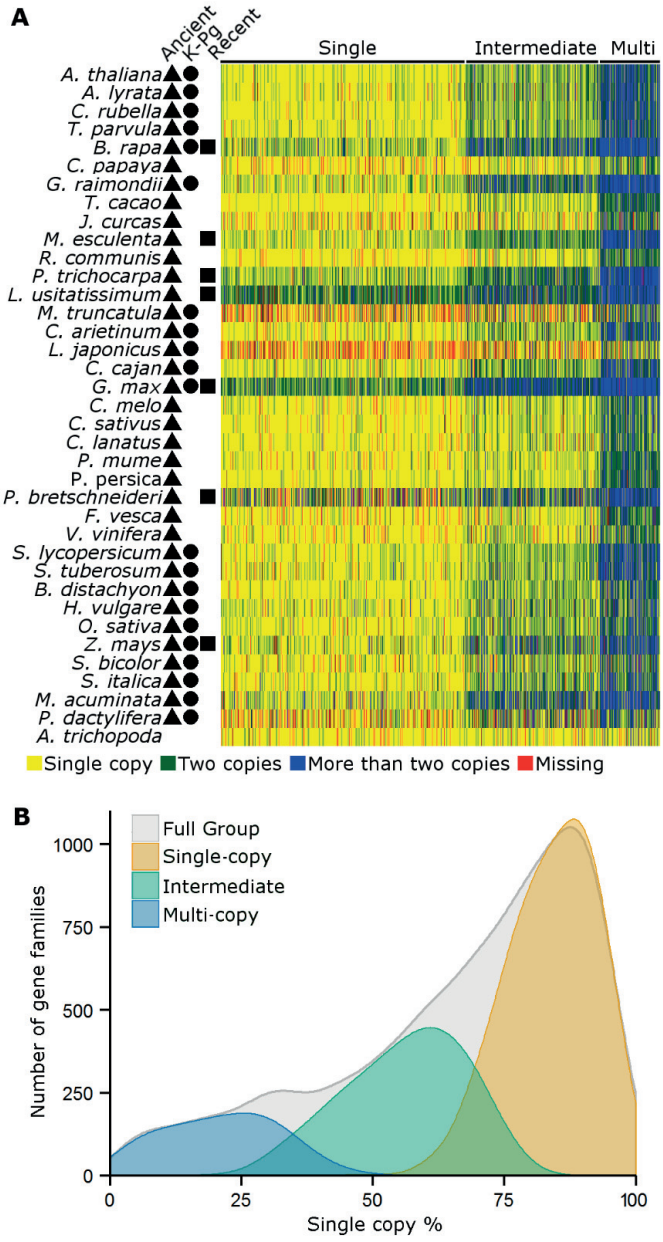


Figure 2.4: ***Core gene families partition into three groups based on clustering of the copy-number profile data. (A)*** *Heatmap of the clustered copy-number profile matrix. Rows represent species and columns represent the core gene families. Gene families (columns) are sorted according to the three different groups obtained by k-means clustering. Symbols indicate for each species whether WGD events that might have contributed to duplicates in the species fall into the 'recent' (rectangle), 'K-Pg boundary' (circle) or 'Ancient' (triangle) category.* ***(B)*** *Single-Copy Percentage distributions for the gene families in each of the three different groups. The 'Cumulative' distribution shows the SCP distribution of all core gene families together (cfr. Figure 2.2).*

Whereas the analyses described above clearly show differences in duplicate retention patterns for the different gene families, it does not provide direct information on the origin of the retained duplicates: e.g.

are duplicates in the Multi-copy group also more ancient than those in the other two groups or is the increased number of species with duplicates in the Multi-copy group mainly due to recent lineage-specific expansions? Therefore, we investigated whether the copy-number patterns observed in Figure 2.4 are related to different ages of retained duplicates in the three groups by using duplication age predictions obtained by GMM of KS-based age distributions and gene tree - species tree reconciliation (see Materials and Methods). The former approach (GMM modelling) provides us with species-specific estimates of duplication ages expressed on continuous time scales (KS -values), whereas the latter approach (reconciliation) gives estimates of the absolute counts of duplication events on a gene family base. Hence, the GMM approach provides multiple estimates of duplicate retention per WGD for events with multiple descendant species, since the modelling is performed in a species-specific manner and as such predictions for the same event are obtained for the species separately. These predictions are not necessarily independent since gene losses following duplication might have predated speciation. However, since KS-values and also their distributions are not always comparable between species [138], the multiple estimates obtained for the same event in different species could not be collapsed. We used the GMM approach to study duplicate retention dynamics over time for gene families in the three different groups, similarly as we did above for the full set of core gene families (Figure 2.3). Overall, when comparing numbers of retained duplicates for the core gene families in function of the WGD ages we observe that gene families in the three different groups differ markedly in their duplicate retention dynamics over time ($p < 9.2e-06$ for all comparisons, Kruskall-Wallis test followed by Dunn's test with Benjamini-Hochberg multiple testing correction) (Figure 2.5A). In particular, we observe higher duplicate retention for all WGD event classes (i.e. for 'Recent', 'K-Pg Boundary' and for 'Ancient' WGD events) for the core gene families in the Multi-copy group, whereas the proportion of core gene families in the Single-copy group with retained duplicates is consistently lower (Figure 2.5A). Next, we used the gene tree – species tree reconciliation approach to obtain absolute counts of predicted duplications and their corresponding ages for all core gene families and used this data to identify group-specific differences in duplicate retention for specific duplication age classes as compared to the full set of core gene families (Figure 2.5B). This shows that gene families in the Single-copy group seem to be specifically biased towards duplicates from the 'Recent' WGDs ($p = p < 2.2e-16$, Fisher's exact test with Bonferroni multiple-testing correction), while duplicates from the 'K-Pg boundary' ($p = p < 2.2e-16$, Fisher's exact test with Bonferroni multiple-testing correction) and 'Ancient' ($p = p < 2.2e-16$, Fisher's exact test with Bonferroni multiple-testing correction) events are underrepresented. Duplicate retention for gene families in the Intermediate group is biased towards the 'K-Pg boundary' events ($p = p < 2.2e-16$, Fisher's exact test with Bonferroni multiple-testing correction). Multi-copy gene families are enriched for duplicates from the 'Ancient' events ($p = p < 2.2e-16$, Fisher's exact test with Bonferroni multiple-testing correction), while showing a deficit in duplications from the 'Recent' events ($p = p < 2.2e-16$, Fisher's exact test with Bonferroni multiple-testing correction). SSDs are underrepresented in the Intermediate group ($p = p < 2.2e-16$, Fisher's exact test with Bonferroni multiple-testing correction), while being overrepresented in the Multi-copy group ($p = p < 2.2e-16$, Fisher's exact test with Bonferroni multiple-testing correction). A comparison of the relative number of duplications obtained for each duplication age class based on gene tree – species tree reconciliation and GMM of KS-based age distributions provide consistent results (Figure S2.8). Despite these differences in duplicate retention for the three groups, all groups have retained more duplicates from the 'Recent' events, followed by the 'K-Pg boundary' and the 'Ancient' events (Figure 2.5A, B).

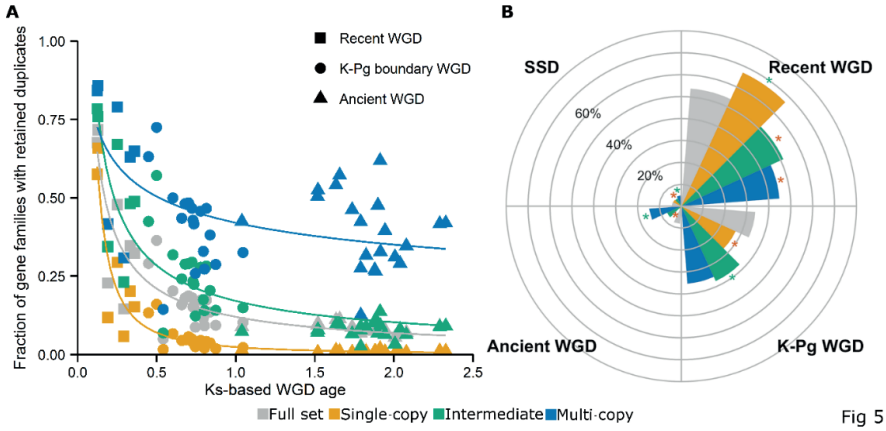


Figure 2.5: ***Analyses of duplication events of the three groups****. (A) For each of the clusters in Figure 2.4, power-law functions were fitted to the corresponding data points representing the fraction of core gene families with retained duplicates following a particular WGD (y-axis) as a function of WGD age (x-axis), as in Figure 2.3 (Chi-squared goodness-of-fit Single-copy group = 0.52, p = 1; Chi-squared goodness-of-fit Intermediate group = 1.38, p = 1; Chi-squared goodness-of-fit Multi-copy group = 1.83, p = 1). The 'Full Set' curve corresponds to the curve represented in Figure 2.3. (B) Polar diagram depicting the fraction of duplication events in each gene family group belonging to either 'Recent', 'K-Pg boundary', 'Ancient' WGDs or 'SSD' events. Here, predicted duplication events were inferred based on gene tree-species tree reconciliation. Green and red asterisks denote statistically significant over- and underrepresentation, respectively, of duplicates of a certain class for a specific group, comparing each time the number of associated duplications for each group with that of the full set (grey bar) by Fisher's exact test. Similar results were obtained by using predicted duplication events inferred using Gaussian mixture modelling of KS-distributions (Figure S2.8).*

2.4.4 The partitioning in different groups is mirrored by gene function

We conducted a GOSlim enrichment analysis of the *A. thaliana* genes in the three different groups, revealing that the three different groups have a remarkably different functional composition (Figure 2.6A). The 'Single-copy' group is enriched for genes that function in organelles (e.g. 'mitochondrion', 'thylakoid' and 'photosynthesis') and that have to do with the maintenance of DNA repair and integrity (e.g. 'DNA metabolic process' and 'nucleobase-containing compound metabolic process'). An independent analysis of 2,090 nuclear-encoded chloroplast-targeted genes taken from The Chloroplast Function Database [139] supported the overrepresentation of genes with chloroplast-associated functions in this particular group ($p = p < 2.2e-16$, Fisher's exact test with Bonferroni multiple-testing correction). No such overrepresentation was found for the 'Intermediate' and 'Multi-copy' groups. The 'Intermediate' group is enriched for genes that are involved in development ('multicellular organism development') and growth and regulation of transcription ('transcription factor activity' and 'chromatin binding'). This last observation was confirmed by an independent analysis of 1,795 putative TFs in *Arabidopsis thaliana* [140], which showed that these genes were clearly overrepresented in the 'Intermediate' group ($p = p < 2.2e-16$, Fisher's exact test with Bonferroni multiple testing correction) while not being enriched for the 'Multi-copy' group and being underrepresented in the 'Single-copy' group. The overrepresentation of regulatory genes in this group, together with the longer retention times for these gene families, suggests that this group mainly consists of dosage-balance sensitive genes [29, 109, 112, 141]. We further investigated this hypothesis by assessing the extent to which genes within this group are involved in protein interactions [103] and the contribution of WGD to duplicate retention for this specific group [29, 103, 105], which represent two characteristics, other than functional

overrepresentation, associated with dosage-balance constraints. First, we observed that *A. thaliana* interacting protein pairs (see Materials and Methods) are indeed most overrepresented in the 'Intermediate' group, yet these results are only borderline significant following multiple testing correction (p = 0.01, randomization test with Bonferroni multiple testing corrections) (Table S2). Second, while all core gene families duplicate preferentially by WGD, the 'Intermediate' group has a higher fraction of WGD-associated duplicates versus SSD-associated duplicates as compared to the 'Single-copy' group (p = p < 2.2e-16, Fisher's exact test with Bonferroni multiple-testing correction) or 'Multi-copy' group (p = p < 2.2e-16, Fisher's exact test with Bonferroni multiple-testing correction), as derived from the gene tree – species tree reconciliation predictions, strengthening our belief that the 'Intermediate' group contains dosage balance-sensitive gene families. Finally, 'Multi-copy' gene families are enriched for genes that appear to be involved in the interaction with the environment ('signal transduction', 'transport' and 'cell wall'), translation, and different metabolic processes ('carbohydrate and protein metabolic process', 'biosynthetic process' and 'catalytic activity').

We also analysed a dataset that describes loss-of-function phenotypes for 2,400 *A. thaliana* genes [142] of which 1,521 are present in the core gene set. Genes within this dataset are placed in four different groups according to their knock-out phenotype. We find that the three core angiosperm groups show markedly different signatures with regards to their classification into LOF phenotype groups (Figure 2.6B). In particular, genes in the 'Single-copy' group are enriched for the 'Essential' category (p = p < 2.2e-16, Fisher's exact test with Bonferroni multiple-testing correction), consisting of genes that are essential for early development and survival. On the other hand, essential genes are underrepresented in the 'Multi-copy' group. This is agreement with recent observations that lethal genes in *Arabidopsis thaliana* usually lack duplicates in this particular genome [143]. Noteworthy, overrepresentation of essential genes in the 'Single-copy' group is not specifically due to the genes involved in DNA integrity within the single-copy set, but also organelle genes are associated with essentiality [142]. The 'Intermediate' set is enriched for genes of the 'Morphological' class (p = 6.96e-05, Fisher's exact test with Bonferroni multiple-testing correction), which contains genes associated with clear morphological phenotypes, involved in reproduction and timing (e.g. flowering time, senescence), in agreement with the strong overrepresentation of developmental genes in this particular group. Finally, the 'Multi-copy' class is overrepresented for genes in the 'Cellular and Biochemical' group (i.e. genes functioning in metabolism, or other biochemical pathways or showing phenotypic effects at the cellular level) (p = 1.14e-06, Fisher's exact test with Bonferroni multiple-testing correction) and 'Conditional' class (p = 6.84e-04, Fisher's exact test with Bonferroni multiple-testing correction) (i.e. genes that respond to biotic and abiotic stress), consistent with GOSlim enrichment results. In summary, both the GOSlim enrichment analysis and the analysis of LOF phenotype data indicate that the separation of core gene families into three different groups according to gene duplicability is mirrored by a separation of the gene families in the space of gene functions.

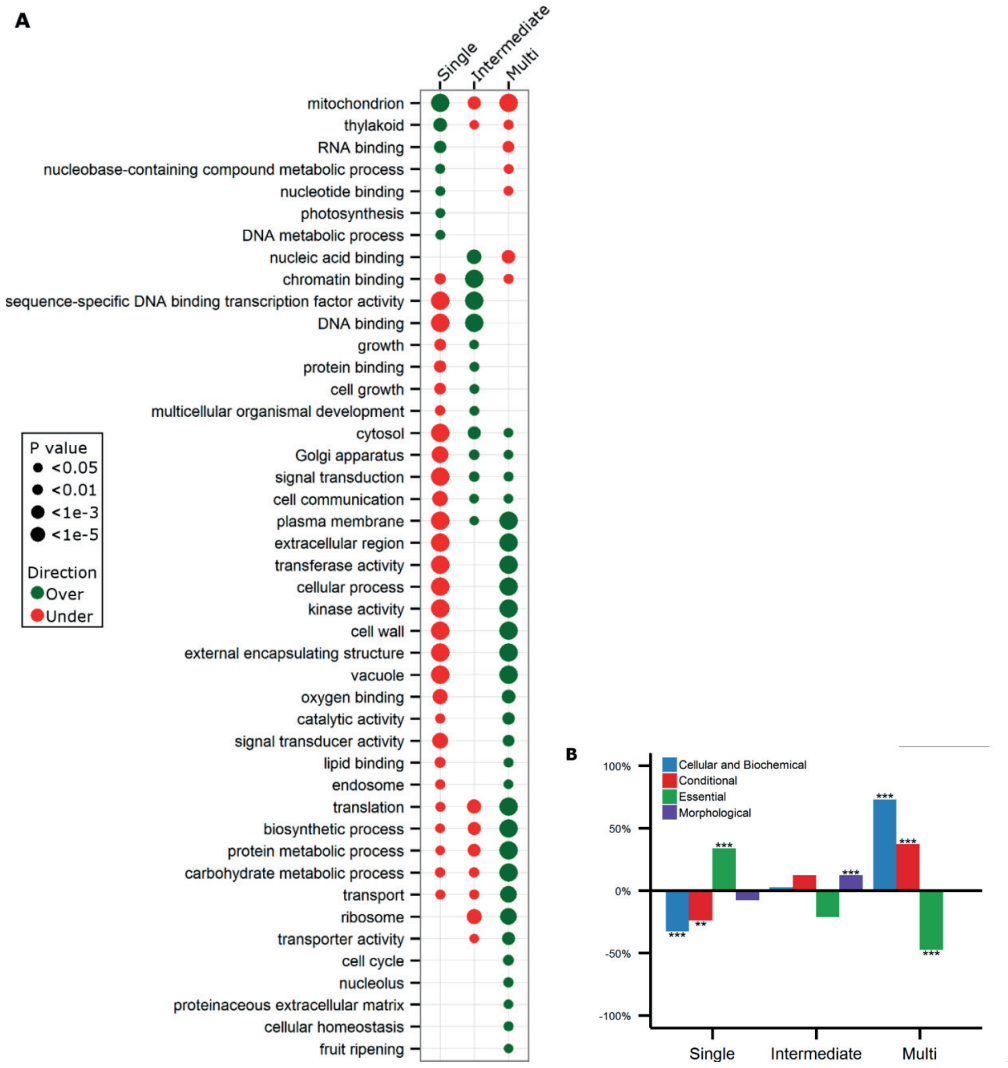


Figure 2.6: ***Functional analyses of the three different groups.*** *(A) GOSlim enrichments and underrepresentation's calculated for the A. thaliana genes in each of the three gene family groups in Figure 4. Dot sizes are representative for the statistical significance of over- (green) or underrepresentation (red). (B) Enrichment analysis of the three gene family groups for knock-out mutant phenotype annotations [52]. Bars represent overrepresentation (positive values) or underrepresentation (negative values) of knock-out phenotypes belonging to any of four functional categories (bar colours). Asterisks denote significance levels as calculated by Fisher's exact test (***: $p < 0.001$, **: $p < 0.05$).*

2.5 DISCUSSION

We assessed duplicate retention patterns for 9,178 core angiosperm gene families (i.e. gene families shared by all angiosperm species) in 37 angiosperm genomes, covering 20 putative WGD events. Assessing the retention of duplicated genes across such a large number of genomes and duplication events allows for replicated tests of gene duplicability, mitigating potential biases due to differences between individual species and WGDs [125-127, 144]. In addition, because of the varied age range of the WGD events in our dataset and the observed large contribution of WGD to the expansion of core gene families, we were able to compare duplicate retention patterns across WGD events of different ages.

We observe that gene duplicability is highly consistent across angiosperm genomes, with over 50% of the core angiosperm genes reverting quickly to single-copy status following duplication, whereas a much smaller set seems to occur in multiple copies throughout. An intermediate group is formed by putative dosage-balance sensitive genes that are maintained in duplicate for prolonged periods of time, but eventually mostly return to single-copy status. By showing that there is a clear distinction between genes that generally occur as a single-copy throughout and genes that show prolonged duplicate retention in the genome or that are retained 'indefinitely' following WGD, we reconcile previous observations on high numbers of single-copy genes shared across multiple angiosperm genomes, despite the many, often nested, WGD events they experienced [33, 121, 123, 124], with observations that duplicates can be retained for long periods following WGD [29, 105]. Previous, smaller-scale comparisons of duplicate retention following WGD in multiple plant species have observed strong differences between species [126, 127]. These differences do most probably exist, yet, by focusing on a large number of species and a large number of WGD events we were able to retrieve dominant and striking patterns of gene duplicability that have remained concealed in smaller-scale comparisons. As our study only focused on core gene families, it is possible that important differences between species result from duplicate retention patterns in gene families that were not considered in this analysis. In addition, while here we showed that the overall duplicate retention tendency seems to be highly consistent across a large number of species and duplication events for the angiosperm core gene families, further detailed cross-species exploration of duplications in both core and non-core angiosperm gene families might reveal other parallelisms in duplicate retention that have remained concealed in this work. For instance, other works have shown that the mode of SSD (primarily tandem versus transposition-duplication) is also preserved cross-taxon for certain gene families [145-147].

We found that gene duplicability is highly associated with gene function, with single-copy genes being biased towards essential genes, functioning in genome integrity pathways and organelles, and multi-copy genes being biased towards functions involved in interactions with the environment. An evaluation of duplicate gene loss and retention patterns following the three successive WGDs in *A. thaliana* uncovered similar correlations between duplicate retention pattern and gene function as the ones observed here [29]. Here, we show that these function-retention patterns can be generalized across a large number of angiosperm genomes and WGD events. In addition, these patterns appear not to be limited to the plant kingdom: in a study focusing on the duplication history of genes across 17 ascomycete genomes, a similar functional separation was observed between genes that generally occur in duplicate and those that are single-copy in most ascomycetes [148]. Likewise, a large-scale analysis of prokaryotic genomes suggested that the number of genes functioning in DNA repair and replication remains relatively constant irrespective of genome size,

whereas the number of TFs, genes involved in signalling and transporter genes, seems to increase with increasing genome size [149, 150]. Consequently, patterns of duplicate retention and loss for core genes in angiosperms and other organisms appear to abide by general function-based rules.

The question remains what causes these specific duplication patterns to occur. Given the overall short half-lives of duplicate genes [28], one could speculate that the observed high fraction of single-copy gene families and a more limited number of multi-copy gene families are caused by a stochastic gene duplication and loss process. We tested this hypothesis and found that stochastic birth-death processes cannot reproduce the observed duplicability distribution, which is heavily skewed towards single-copy gene families. In addition, the observed overall consistency of patterns across genomes and across large-scale duplication events and the functional enrichments observed for the various duplicability classes of gene families argue against such a random scenario. Considering the strong association with gene function, a possibility is that gene function directly or indirectly constrains gene duplicability. The observed patterns of gene duplicability are indeed consistent with the idea of the existence of a conserved core, that needs to remain untouched ('Single-copy' group), and the existence of processes that are more amenable to modifications and that might be responsible for adaptations to new environments and the evolution of distinct morphological features ('Multi-copy' group) [151]. Gene duplication in itself can indeed modulate gene function in a negative way and as such impact core gene function, by for instance increasing absolute gene dosage of genes with strict gene expression constraints [152], through the accumulation of mutations in duplicate copies with potential pleiotropic negative effects on wild-type fitness [33, 153-155] or potential cytotoxic effects (e.g. protein misfolding) [156]. As a result, duplicates of genes sensitive to these processes might be eradicated quickly, also after WGD. On the other hand, repeated biased retention of certain duplicates for long periods of time ('Intermediate' group) or indefinitely ('Multi-copy' group) suggests a mechanism of duplicate retention other than sub-/neofunctionalization, which are in general assumed to be slow processes [157] and would not be expected to lead to repeated biased retention. Considering the primary role of WGD in duplicate retention of the core genes and the specific association of gene functions enriched in the ‘Intermediate’ and ‘Multi-copy’ group with previously defined putative dosage-balance sensitive genes [29, 134], we hypothesize that dosage-balance constraints may have contributed to the prolonged retention of duplicate genes in these sets. Prolonged retention of duplicate genes, accompanied by gradual circumvention of dosage balance constraints, may increase the possibility that duplicate genes diversify and get permanently preserved [111, 144]. Alternatively, duplicate genes could also be permanently retained through absolute dosage constraints replacing over time the relative dosage-balance constraints responsible for initial duplicate retention [128, 158]. In our results, the ‘Intermediate’ group of gene families exhibits the hallmarks of dosage-balance constraints that wear off over time, leading to prolonged preservation and ultimately loss of duplicates. A subset of genes in the ‘Multi-copy’ group may also have been retained initially because of dosage-balance constraints and, in this instance, preserved indefinitely through other mechanisms; in particular transporters, signalling transducers and cell communication genes have been reported earlier as potentially dosage balance-sensitive [29, 105]. On the other hand, the ‘Multi-copy’ set of gene families is also enriched in ‘environmentally responsive’ genes. Consequently, their repeated and biased retention following WGD might be a consequence of an increased adaptive advantage of polyploidy under environmental stress. Indeed, increasing evidence suggests that polyploids show wider environmental tolerance and higher levels of phenotypic plasticity than diploids [22, 159-165]. In particular transporters and metabolic genes, enriched in

the 'Multi-copy' class, have been identified before as putative driver genes explaining the increased tolerance of polyploids for environmental stress [160, 162, 166-169]. Despite the strong correlation between gene duplicability and gene function observed here, it remains to be further investigated which evolutionary mechanisms are responsible for the observed strong bias in duplicate retention patterns, and it remains to be established whether gene function directly influences gene duplicability or whether biased gene retention could be a by-product of other evolutionary phenomena instead, such as for instance the preservation of intermolecular interactions (dosage balance) or sequence constraints related to high levels of gene expression [96, 170]. In particular, since network structure is often believed to constrain protein evolution and to underlie complex phenotypic traits, future work into this direction might benefit from investigating gene duplicability in a network context (e.g. [128, 158, 171-173]).

2.6 MATERIALS AND METHODS

2.6.1 Genome Data

We employed protein-coding genes from 37 fully sequenced angiosperm genomes, 35 of which were used in [108]. Protein-coding sequences for Amborella trichopoda [137] and Capsella rubella [174] were retrieved from the Amborella Genome Database (http://www.amborella.org/) and Phytozome V10, respectively.

2.6.2 Gene Family Prediction

OrthoMCL

We identified gene families based on protein sequence similarities by OrthoMCL [175]. After all-against-all BLASTP searches, OrthoMCL was used to group proteins with high sequence similarity into gene families. An important parameter of OrthoMCL is the inflation parameter, which controls cluster tightness. We calculated gene families for different inflation parameter values (i.e. 1.5, 2.0, 2.5, and 3.0) to assess its influence, and observed large variations in the number of gene families detected and their overall size. We decided to use the inflation parameter that gives on average the largest gene families (i.e. 1.5), since the gene families are further processed by phylogenetic tree construction (and split up if necessary, see below). As such we obtained 69,133 multi-gene families.

Species tree construction

A species tree was constructed from a concatenated multiple sequence alignment inferred from 107 gene families that are present in all of the 37 angiosperm species and contain no more than 40 genes in total. The genes within these 107 gene families are on average longer than 150 amino acid residues. If a species had paralogs in a gene family, we only kept the paralog with the most orthologous hits in the gene family in the intermediate OrthoMCL results file. We used MUSCLE (3.8.31) [176] with default parameters to perform multiple sequence alignments for each gene family based on the amino acid sequences. We then used trimal (1.4) to remove low quality regions of the alignments based on an automatically selected threshold (-strictplus), which depends on a distribution of residue similarity inferred from multiple sequence alignment for each gene family [177]. Multiple sequence alignments of amino acid sequences were back-translated into alignments of codon sequences and were concatenated one by one into an integrated alignment. In the end, we obtained an alignment of 36,631 codons with 109,893 nucleotide sites.

To construct the species tree, we used CodonPhyML (1.0) [178] under three different codon models that differ in their instantaneous substitution rates between codons, being the Muse and Gaut (MG) model [179], the Goldman and Yang (GY) model [180] and the YAP model [181]. The stationary frequency of codons and the transition-transversion ratio were estimated by maximum likelihood. The different ratios of nonsynonymous to synonymous substitution rate over the sequence alignment were drawn from a discrete gamma distribution with three, four, or five classes. The parameters α and β of the gamma distribution were optimized by maximum likelihood. An initial tree was built using the BioNJ algorithm, based on the empirical model ECMK07. CodonPhyML then employs Nearest Neighbour Interchange (NNI) and Subtree Pruning and Regrafting (SPR) to optimize the tree topology. Branch lengths and model parameters are also fully optimized during this process.

Based on the different codon models and parameters described above, we obtained nine phylogenetic trees with identical topology but with slightly different branch lengths. The branch lengths of the different trees have no effects on the phylogenetic placement of WGDs. We used the Akaike Information Criterion (AIC) to compare likelihoods for the different trees and selected the tree with the lowest AIC tree as the species tree in this study. This tree corresponds to the tree inferred under the MG model with five classes for ω.

We calculated bootstrap support values for all branches of the species tree by obtaining 100 bootstrap samples for the concatenated multiple sequence alignment and running CodonPhyML on each bootstrapped alignment using the same model and parameter settings as chosen for the species tree. The bootstrap values were added on each branch of the species tree by RAxML [182]. As an alternative support measure to the bootstrap we assessed the degree of congruence between the species tree topology and the topology of the 107 gene trees, also obtained using CodonPhyML with the same parameter settings, for the gene families used for species tree construction. Specifically, using RAxML, we calculated two measures: (1) internode certainty (IC) and (2) IC All (ICA) that evaluate the support for an internode in the species tree by considering its frequency in the set of 107 gene trees [183, 184]. An Internode Certainty value of one means that none of the gene tree topologies conflict with the species tree topology, whereas a value close to zero for internodes suggests that there is another possible bipartition that occurs with almost equal frequency to the inferred one. In the end, the species tree was rooted on the branch of the basal angiosperm species Amborella trichopoda and was visualized by FigTree (http://tree.bio.ed.ac.uk/software/figtree/). This obtained species tree is largely consistent with the APGIII tree [185].

Gene tree construction and reconciliation

Next, we implemented a pipeline to automatically construct phylogenetic trees for all 69,133 gene families and to test whether these trees could be traced back to a single angiosperm ancestral gene. We first removed 253 gene families with more than 200 genes because of the enormous computational resources required by large gene families. Then we built maximum likelihood phylogenetic trees for each of the remaining gene families with more than two genes. Multiple sequence alignments based on protein sequences were produced using MUSCLE with default settings [176] and were further trimmed by trimal in a heuristic automated approach (-automated1) [177]. The processed multiple sequence alignments were fed into PhyML 3.0 [186] using the LG model with the equilibrium frequencies defined in the substitution model. The best trees produced from either Nearest Neighbor Interchange or Subtree Pruning and Regrafting were

retained as maximum likelihood gene trees. To obtain branch support values for the gene trees, we used the SH-like approximate Likelihood-Ratio Test [187] instead of traditional bootstrap values because of its speed.

For 28,946 gene families with at least four genes from at least two different species we used gene tree-species tree reconciliation [188] to root the gene trees and to obtain estimates of duplication and speciation events along the gene tree. For the remaining 39,934 gene trees, prediction of duplication and speciation events is trivial (see below). Since the reconciliation process is error prone [189-191] and depends on the quality of the gene tree, species tree and the parameter settings of the reconciliation method we implemented a pipeline to mitigate these problems as much as possible: (1) Since PhyML does not explore the entire search space of possible tree topologies, we investigated whether alternative tree topologies with improved reconciliation duplication/loss costs, obtained by branch rearrangements of the original gene trees in the reconciliation strep (see below), had an increased likelihood under the multiple sequence alignment than the gene tree produced by PhyML. As such we obtained a reconciled gene tree that is maximally supported by both the reconciliation criterion (in this instance duplication/loss cost) and the multiple sequence alignment as described in Wu, Rasmussen [190]; (2) To deal with the problem of reconciliation solutions being dependent on the parameter settings we performed the reconciliation with a range of different parameter settings and we also considered multiple possible optimal reconciliations under the same parameter settings, if available. Since duplication/speciation events that were predicted for multiple parameter settings are assumed to be more reliable [191], we built a majority-rule consensus reconciliation in which we only retained duplication/speciation events supported by at least 50% of the reconciliations.

If a duplication event was predicted at the Angiosperm-associated node, we split the phylogenetic tree into two subtrees (and hence also two associated gene families), ensuring that each subtree traced back to a single ancestral Angiosperm gene. With this procedure, we obtained 11,131 gene families with gene trees tracing back to an angiosperm ancestral gene. From this set we removed the gene families that did not have gene copies for at least 32 out of 37 species (Figure S2.1), ending up with a final set of 9,178 core gene families.

For the remaining 39,934 gene families (i.e. gene families with at least two species but no more than three genes or gene families that are only present in one species), we inferred duplication events by simply applying the following rules (see Figure S 2.9). For gene families with only one species, after mid-point rerooting of the gene tree, each node in the tree represents a duplication node. For gene families with two genes, after mid-point rerooting of the gene tree, nodes were annotated as duplication nodes if the two genes were from the same species. For gene families with three genes we used the topology of the gene tree to infer the duplication events.

2.6.3 KS-based age distributions

KS-based estimation of timing of duplication

Estimates of KS-values were obtained for all paralogous pairs associated with the predicted duplication events inferred by the gene tree/species tree reconciliation process. For cases where there are multiple possible pairs for a predicted duplication event, we calculated KS-values for all possible gene pairs and selected the gene pair with the smallest KS-value to represent the timing of the duplication event. For each

paralogous gene pair we aligned the protein coding sequences using ClustalW [192] using parameter recommendations from [193]. PAL2NAL [194] was used to back-translate the aligned amino acids into corresponding codons without gaps. Then codeml [180] from PAML [195, 196] was used to obtain KS-values for each gene pair using the GY model with stationary codon frequencies empirically estimated by the F3x4 model.

Gaussian Mixture Modelling of KS-based age distributions

For each species in our dataset we fitted Gaussian mixtures to age distributions inferred from KS-values [133-135], using the R-package 'mixtools'. We ignored KS-values that exceeded 5.0. First, we determined for each age distribution the number of components (k) using the 'boot.comp' function. Specifically, we performed parametric bootstraps with 1000 bootstrap realizations of the likelihood ratio statistic for testing the null hypothesis of a k-component fit versus the alternative hypothesis of a (k+1)-component fit. For this test, a significance level of 0.01 was used. For each age distribution, we tested the presence of one to 6 components. The number of components determined in this first step was used to fit a mixture of Gaussian models to the KS distribution, using the 'normalmixEM' function with the following parameters: k=k, maxit = 1e30, maxrestarts = 1e3, epsilon = 1e-50. We manually curated the obtained peaks, only further focusing on solid WGD peaks (Figure S 2.10). Dispersed background peaks with mean μ >3 and model peaks with obvious misfits to the data were ignored for the purpose of duplication assignment. We assume that each remaining peak corresponds to a WGD event, except for the first peak, which likely consists of recent SSDs [29]. A duplication was assigned to the peak that showed the highest probability density at the KS value obtained for its representative paralog pair [29]. For each WGD, we obtain an associated estimate of the number of gene families with retained duplicates as the ratio of the number of core gene families with duplicates for that event to the total number of core gene families. Each peak was characterized by an age (expressed in KS -values) that corresponded to the mean (μ) of the Gaussian mixture component (see Table S1 for detailed peak information). To assess duplicate retention in function of time since duplication we plotted duplicate retention associated with a certain WGD (y) in function of the predicted age of that event (x). We then fitted exponential and power-law functions to these data. Both functions have previously been used to describe the relationship between duplicate retention and time since duplication [28, 29]. In all instances, the power-law fit was preferred over the exponential fit based on the Chi-squared goodness-of-fit measure (Figure S2.11, Table S3).

2.6.4 Evolution of gene families under a stochastic birth-death null model

The null model

The null hypothesis describes the evolution of gene families along the phylogeny as a random birth-death (BD) process with equal rates of SSD gene duplication and loss per evolutionary time unit (unit branch length), λ, as proposed by Bailey [130]. Since WGDs violate the assumption of independency of duplication events in Bailey's BD model [130], we have placed these events as separate nodes on the branches of the species tree, similar to the strategy employed by Rabier, Ta [131]. At WGD nodes, all gene family members are instantaneously duplicated (or triplicated, depending on the nature of the polyploidy event). As in the model of Rabier, Ta [131], we assume that a given fraction of duplicates is lost very quickly after WGD, represented

by an immediate loss rate parameter q in our model. The remaining WGD duplicates are lost over time at a loss rate λ, the same as for SSD duplicates. A full description of the model will be published elsewhere.

Our purpose is to use this BD model to generate gene counts at the leaves of the species tree for a number of simulated gene families and compare the Single Copy Percentage (SCP) distribution of these simulated families to the SCP distribution observed for the core gene families. In each run, we simulated gene counts under the random BD model for 9,178 gene families, corresponding to the number of families in the core set. We performed 1,000 such runs and estimated the SCP null distribution as a kernel density function over the 9,178 x 1000 simulations.

For each simulated gene family, we sample a value for λ and q from predefined distributions (see below), and we assume that the root size - the gene count at the root of the species tree - is equal to 1. We start at the root and generate a gene count for each of the child nodes of the root through an MCMC process that samples a child node size from the node size probability distribution function described in the BD model [130]; 5000 MCMC steps were used as burn-in to guarantee MCMC convergence to the stationary BD probability distribution. The same procedure is used for any further progeny node up to the leaf nodes, each time starting from the previously generated gene count at its parent node. At WGD nodes, the node size is multiplied after node size sampling with 1+ d.(1-q) to mimic the WGD effect, with d=1 for duplications and d=2 for triplications. In our simulations, we imposed the limitation of generating at least 32 non-zero gene counts at the leaves of the species tree, to be consistent with the fact that the core gene families studied were required to be present in at least 32 out of 37 species.

The q value to be used for a given duplicate birth-death simulation is uniformly sampled from the range [0-1], with 0 being complete retention and 1 complete loss of duplicates immediately after WGD (q is assumed to be the same for all WGDs across the tree, i.e. it is assumed to be a property of the gene family). The λ-value to be used for a given simulation is sampled from a normal distribution with mean λ_{av} = 0.53 and standard deviation σ = 0.156. The rationale for sampling birth rates from this specific distribution is the following. We assume that the average duplication rate per gene, λ_{av}, is approximately equal to the average synonymous substitution rate per synonymous site [135], i.e.:

$$\lambda_{av} = \frac{\text{average \#duplications/gene}}{t \text{ time unit}} \approx \frac{\text{average \#synonymous substitutions/syn.site}}{t \text{ time unit}} = \frac{\text{average } K_S}{t \text{ time unit}} \quad (1)$$

where 't time unit' stands for the evolutionary time unit used in the species tree (where branch lengths are expressed in terms of the number of substitutions per codon t), i.e. the evolutionary time needed to obtain one substitution per codon on average (unit branch length t=1). To assess approximately how many synonymous substitutions per synonymous site (KS) are expected to occur per t time unit in an average plant DNA sequence, we inferred an average relationship between t and KS from the following formula for the number of substitutions per codon t in a given sequence [197]:

$$t = \frac{(K_S \times S) + (K_N \times N)}{\frac{S+N}{3}} \quad (2)$$

with S and N, the number of synonymous and non-synonymous sites in the sequence and K_S and K_N the number of synonymous and non-synonymous substitutions per (non)-synonymous site, respectively. Equation (2) can be rewritten as:

$$t = 3\,K_S \times\ (1 + \frac{\omega - 1}{\frac{S}{N} + 1})\ (3)$$

with $\omega = K_N/K_S$ the ratio of non-synonymous substitutions per non-synonymous site to synonymous substitutions per synonymous site, and S/N the ratio of synonymous sites to non-synonymous sites in a sequence. For both ω and S/N, we substitute genome-wide average estimates to obtain an approximate relationship between t and KS for an average sequence evolving under average selective pressure. Taking S/N = 0.345 for the average codon [198], and taking an ω value of 0.5 on average (as observed for Arabidopsis duplicates in the K_S range [0,1] [133]), the following estimate of t as a function of K_S is obtained for the average plant DNA sequence:

$$t \approx 1.884\,K_S\ (4)$$

In other words, in one t time unit, $1/1.884 \approx 0.53$ synonymous substitutions are estimated to have accumulated per synonymous site on average. We use this estimate in equation (1) to obtain an estimate of the average duplication rate per gene $\lambda_{av} = 0.53/\text{gene}/(t\ \text{time unit})$. To assess how this λ_{av} estimate compares to literature estimates of duplication rates expressed per gene per million years, we used the average duplicate K_S and absolute age estimates for fairly recent WGDs (0 < K_S < 1, in the range where K_S estimates are reliable) reported by Vanneste, Baele [108] to convert the resulting estimate $\lambda_{av} = 0.53/\text{gene}/(t\ \text{time unit}) = 1/\text{gene}/(K_S\ \text{time unit})$ to an estimate of the duplication rate expressed per million years (here, one K_S time unit is the evolutionary time it takes to obtain $K_S = 1$ on average, which corresponds to $1/0.53 \approx 1.884$ t time units according to equation (4)). By dividing the average WGD duplicate pair K_S estimates by twice the absolute WGD age estimates reported in [108] (note that the evolutionary time elapsed between WGD duplicates in My is twice the age of the WGD), and averaging over all WGDs, we get a K_S /My conversion factor of 0.00585, giving $\lambda_{av} = 0.00585/\text{gene/My}$, which is reasonably comparable to earlier estimates of duplications/gene/My across species [135, 199]. With the average duplication rate λ_{av} in our tree estimated at 0.53/gene/(t time unit), we defined a λ-distribution around this value with standard deviation 0.156, so that more than 99% of the probability mass lies within the λ interval [0-1]. Qualitatively similar results were obtained with other λ_{av} values and λ-distribution shapes (results not shown).

Dating WGDs

To run the simulations described above, WGD events need to be added to the phylogenetic tree as new nodes with known branch lengths in terms of t, the number of substitutions per codon. To this end, for each of the WGDs, we averaged the t estimates for all (predicted) homeologs for which the K_S estimates fall within the WGD K_S range described in Vanneste, Baele [108]. t and K_S estimates for all homeolog pairs were obtained using codeml [180] as described in [108]. As we repeated this procedure for each species separately (except for Capsella rubella and Amborella trichopoda, which were not analysed in Vanneste, Baele [108]), multiple t estimates were obtained for shared WGDs. In this case, we used the average species-specific t-estimates to position a given shared WGD on the tree.

All of the resulting WGD estimates were positioned on the species phylogeny in a manner consistent with their taxonomic positioning reported earlier [108, 129], except for the most recent WGDs in *Gossypium raimondi* and *Zea mays*, which were inferred by our t-estimation protocol to be positioned on older branches

than the accepted ones, likely because of t and K_S estimation and averaging inaccuracies. In these cases, we positioned the WGD in the beginning of the branch reported in literature.

2.6.5 Clustering of the copy-number profile matrix

To determine gene family-specific differences in duplicate retention, the gene family data was transformed into a count matrix, in which elements represent the number of gene copies for a certain gene family (columns) in a certain species (rows). To reduce the influence of outliers (families with lots of genes), we only used gene families with maximum three gene copies per species. We clustered this matrix in the direction of the gene families using ConsensusClusterPlus, which incorporates a subsampling approach to infer cluster number and cluster confidence [136, 200]. This R implemented package was run using the following options: maxK = 8, reps=100, pItem=0.8, pFeature=1, K-means, inner linkage=average, final linkage=average, distance=Pearson. A solution with three clusters was found to be optimal according to the built-in cluster stability criterion (Figure S2.6) [136]

2.6.6 Functional data

PPI data *A. thaliana*

A compendium of protein-protein interactions in *A. thaliana* was constructed combining the following sources, BioGRID 3.2.110 [201], CORNET (only experimentally validated interactions) [202], STRINGv9.1 (only category Binding) [203], EVEX (only category binding) [204] and a TAP dataset assembled from literature [205-219]. After removing redundancy and self-interactions this lead to a set with a total of 46,113 interactions between 9,813 proteins.

Enrichment of PPI, LOF, Chloroplast genes and Transcription factors

The Fisher's exact test was used to calculate if a class is overrepresented in a given set of genes. In order to test whether there are more protein interactions within a group than between a group, 1000 randomized interaction networks with the same degree distribution were constructed. For each group of genes, a z-score was obtained by comparing the number of protein interactions within the group based on the extant PPI network with the distribution of within-group interaction counts observed in the randomized networks. Z-scores were then converted into one-tailed p-values.

Functional enrichment analysis

The BINGO 2.44 Cytoscape plugin [220] was used to calculate functional enrichment values for the set of *A. thaliana* genes. We used a p-value threshold of 0.05 and p-values were corrected for multiple testing using the Benjamini and Hochberg method [221].

2.7 SUPPLEMENTARY INFORMATION

This section contains selected segments of supplementary methods, figures and tables most relevant to this dissertation. The full supplemental information can be found on

- Supplemental Figures and Tables

http://www.plantcell.org/content/plantcell/suppl/2016/01/07/tpc.15.00877.DC1/TPC2015-00877-LSBR1_Supplemental_Data_pdf.pdf

- Supplemental Dataset 1

http://www.plantcell.org/highwire/filestream/4486/field_highwire_adjunct_files/0/TPC2015-00877-LSBR1_Supplemental_Data_Set_1.txt

- Supplemental Dataset 2

http://www.plantcell.org/highwire/filestream/4486/field_highwire_adjunct_files/1/TPC2015-00877-LSBR1_Supplemental_Data_Set_2.xlsx

2.7.1 Supplementary figures

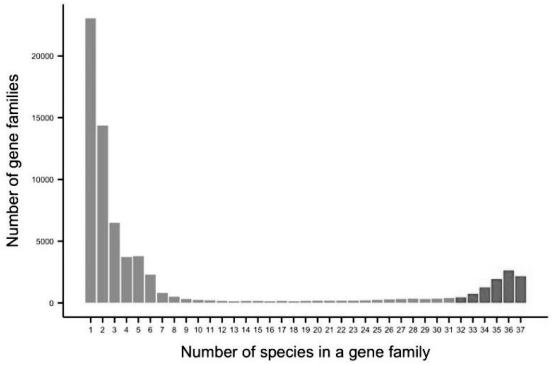


Figure S2.1: ***Motivation for the 32 out of 37 species cut-off to define core gene families****. To distinguish core from non-core gene families we assessed the distribution of the number of species in each gene family based on all 69,542 gene families obtained by reconciliation. This distribution is U-shaped, suggesting a large number of gene families that are species- or lineage-specific (left side of the distribution) and also an excess of gene families present in the large majority of angiosperm species (right side of the distribution). Based on this distribution we decided to consider all gene families containing genes from at least 32 species as being 'core gene families'. As such we account for a limited number of putative missing orthologs from core gene families due to for instance errors in genome annotation, gene family construction errors or the presence of incomplete genomes.*

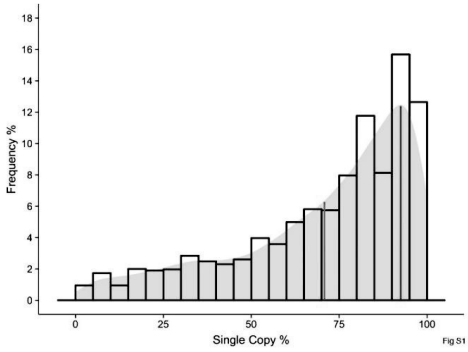


Figure S2.2: ***The distribution of Single-Copy Percentages (SCPs) for all core gene families****, with SCPs calculated upon removing the highly duplicated genomes of Glycine max, Linum usitatissimum, Brassica rapa, and Zea mays. This distribution has a mode of 92% and a mean of 70.8%.*

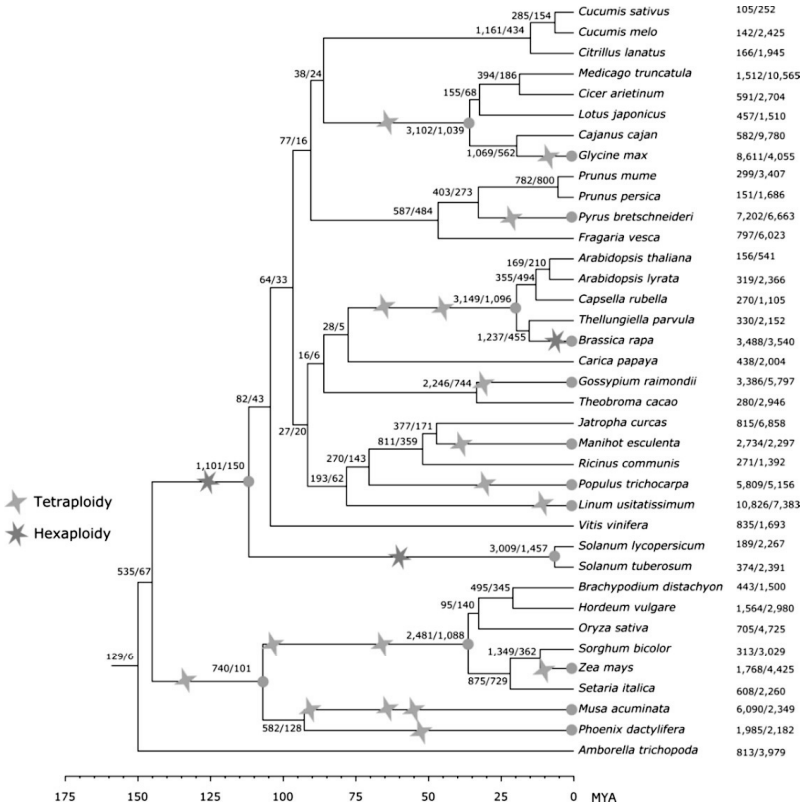


Figure S2.3: ***Classification of species tree nodes as SSD or WGD****. On the species tree, nodes with WGDs on their parent branches were considered as WGD nodes (orange dots), while the rest of the nodes were considered as SSD nodes. Next to each node are the number of duplication events predicted by gene tree-species tree reconciliation for both core and non-core gene families (core/non-core). There are in total 93,942 predicted duplication events in core gene families and 140,786 duplication events in non-core gene families.*

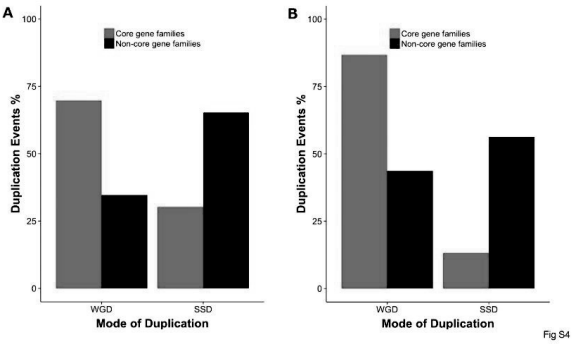


Figure S2.4: ***Core gene families mainly duplicate through WGD****. Bar plots represent the fraction of duplication events, summed over all gene families, attributed to WGD or SSD in core and non-core gene families. Panel (A) represents results obtained from all nodes in the species tree in (Figure S2) and shows that for core genes families, as compared to non-core gene families, the presence of duplicates seems to be biased towards WGD-associated gene duplication (p < 2.2e-16, Fisher's exact test). In panel (B) we assessed the possibility that these observations might be caused by an overrepresentation of WGD-associated nodes in the species tree for core gene families as opposed to non-core gene families: since core gene families cover by definition a larger number of species, some of the more ancient WGD events that are shared by many species will only be represented by core gene families. Hence, we repeated this analysis by only*

considering nodes from the species tree that are also ubiquitously present in non-core gene families (top 10 of the nodes) and came to the same conclusion (p < 2.2e-16, Fisher's exact test).

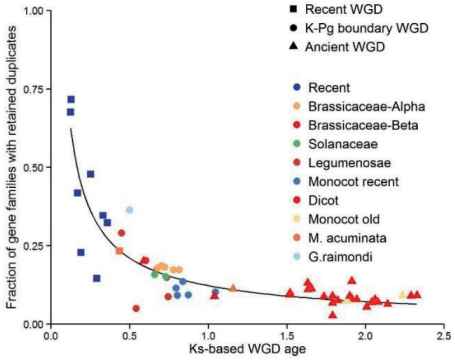


Figure S2.5: ***Duplicate gene retention in function of time since WGD****. Each dot represents the fraction of core gene families with retained duplicates following a specific WGD (y-axis), as a function of WGD age, expressed in K_S-units (x-axis). The timing of the WGD events and the particular gene families that retained duplicates following a specific WGD event were inferred by fitting Gaussian mixture models to K_S-age distributions for all 37 species separately (see Materials and Methods). This figure is related to Figure 3, but here all WGD peak callings were included. Since the Dicot and Brassicaceae-Beta peaks cannot be distinguished from each other they are denoted by the same colour. Additional information on all the peaks is provided in the full supplemental data.*

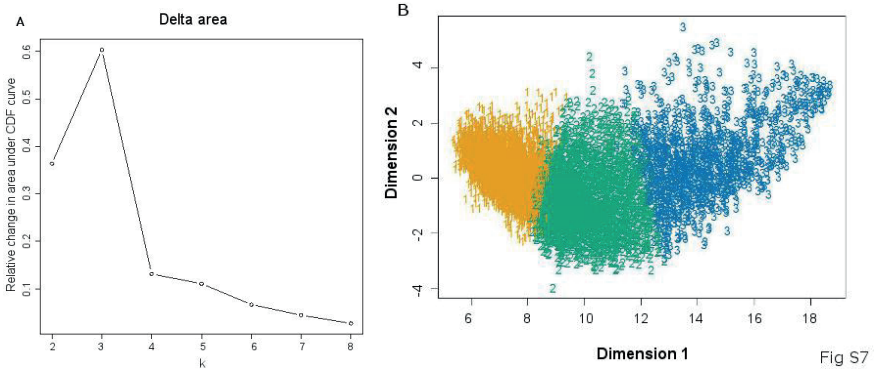


Figure S2.6: ***Criteria that we used to choose the optimal number of clusters for k-means clustering of the copy-number matrix.*** *(A) We used the Delta Area Plot from the ConsensusClusterPlus R-package to select the optimal number of clusters. The results of 1000 clustering runs, each time on subsampled matrices, are summarized into a consensus matrix, whose values represent the proportion of clustering runs in which two items (i.e. gene families) are grouped together. Hence, values in this matrix are between 0 and 1 (1 = always clustered together). The Delta Area Plot assesses the 'cleanness' of this consensus matrix: if all clustering runs agree on the same solution than this matrix only consists of 0's and 1's (bimodal distribution). To determine the optimal numbers of clusters the largest changes in these consensus values are detected by calculating the change in the area under the Cumulative Distribution of consensus values for increasing cluster number (Monti et al. 2003). The 'Delta area' represents this change, with k corresponding to cluster number. (B) Corresponding multidimensional scaling plot of the copy-number matrix, with data points coloured according to cluster membership.*

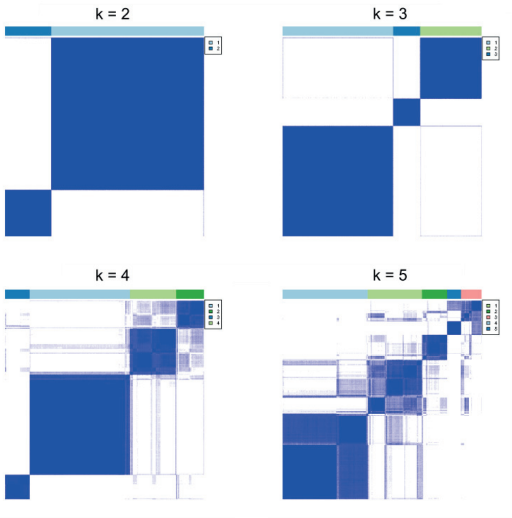


Figure S2.7: ***Consensus matrices obtained for different number of clusters k****. The consensus matrix represents the number of times that two gene families belonged to the same cluster over 1,000 clustering runs of the subsampled copy-number matrix. The values within this matrix range from 0 (gene families were never grouped into the same cluster; white in this figure) to 1 (gene families were always grouped into the same cluster; blue in this figure). Here results are shown for k = 2-5 clusters. Colour bars on top of the visualized consensus matrix indicate cluster assignments.*

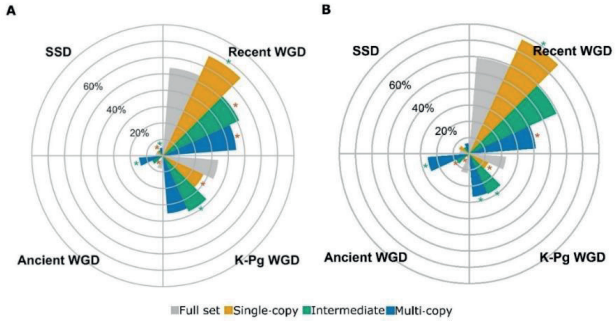


Figure S2.8: ***Polar diagrams depicting the fraction of duplication events in each gene family group belonging to either the 'Recent', 'K-Pg Boundary', 'Ancient' or 'SSD' duplication classes.*** *(A) Represents predictions of duplication timing for all core gene families, obtained by using gene tree – species tree reconciliation. This Figure is the same as Figure 5B. In contrast to GMM (see panel B), which provides estimates of the ages of the duplication events for each species separately, here estimates of the duplication age is based on a gene family basis and hence no averaging over species is necessary. To obtain the bar plots we normalised the absolute counts of duplication events for each node in the species tree with the number of nodes in the species tree of that duplication class, correcting for the fact that there are for instance more nodes associated to the 'SSD' duplication class. Significance values are indicated by asterisks (green = overrepresentation, red = underrepresentation) and were calculated based on the absolute counts of predicted duplications of each class, using the Fisher's exact test with Bonferroni multiple-testing correction. (B) Represents predictions of duplication timing for all core gene families based on GMM of KS-based species-specific age distributions. We classified each duplicate pair to a certain duplication class depending on the KS-peak it belonged to (see Table S1). The bars in the Figures represent averages, obtained from averaging over the number of duplications assigned to a certain class for all species. Statistical significant over- and underrepresentation's were calculated based on the Wilcoxon-rank-sum test and are denoted by asterisks.*

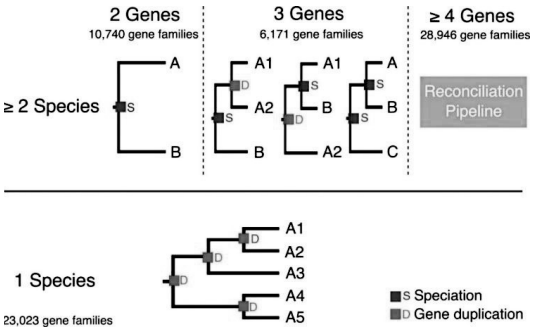


Figure S 2.9: ***Explanation of how duplications were inferred for gene families with at least two species but no more than three genes or gene families that are only present in one species****. For gene families with two genes in two species (10,740 gene families), the node connecting both genes is assumed to be a speciation node. For gene families with three genes (6,171 gene families), we mid-point rerooted the gene tree and distinguished between three possible scenarios. If the three genes come from two species, the duplication occurred either in one species or in the common ancestor of the two species, depending on the topology of the gene tree. If the three genes come from three species, we assume that no duplications have occurred in the history of the gene family (most parsimonious scenario). For gene families that only cover one species (23,023) but with two genes or more, e.g. five genes in the figure, we mid-point rerooted the gene tree and considered all nodes in the tree to be duplication nodes. For the remaining 28,946 gene families with at least four genes (including all core gene families) duplications were inferred using the reconciliation pipelines as described in Supplemental Methods.*

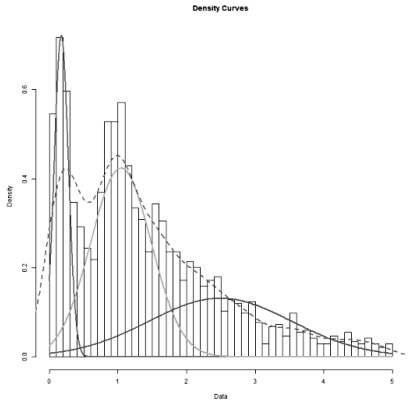


Figure S 2.10: ***Gaussian mixture models were fit to the KS-distribution of each species.*** *Peaks were considered solid if they had a good visual fit with the density line (dashed purple line) and the KS-histogram and had a lower than 3. Flat peaks, e.g. peaks which span the whole KS - distribution, where also removed. The annotation of the peaks was done using known literature [108]. The figure shows the KS -distribution for Sorghum bicolor. The red and green peaks have a good fit to the density line whereas the flat blue peak shows no correspondence to density line and spans the whole KS - distribution.*

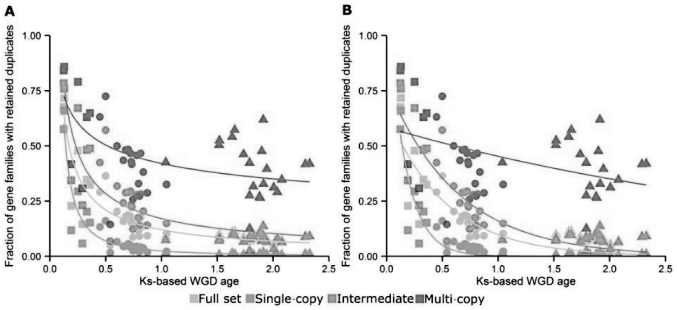


Figure S2.11: ***Comparison of (A) power-law fit and (B) exponential fit to the data obtained from the Gaussian Mixture Modelling of KS-based age distributions.*** *The power-law shows consistently a better fit than the exponential, as assessed by Chi-squared Goodness-Of-Fit test (see Table S3).*

2.7.2 Supplemental tables

Table S2.1: ***Comparison of the numbers of interacting protein pairs in each group to those obtained from randomized networks.***

	Number of PPIs within group	Average number of PPIs within group for 1000 randomized networks	Z-score	P-value enrichment of PPI vs random (one-sided test)	P-value with multiple-testing correction (Bonferroni)
Full	15949	15949			
Single-copy	2550	2813.012	-1.005	0.84	1
Intermediate	2277	1740.331	2.710	0.0034	0.010
Multi-copy	1034	990.558	0.322	0.374	1

Table S2.2: ***Comparison of the power-law and the exponential fit.***

	Chi square goodness-of-fit (p-value)	
	Power-law	Exponential
Full	0.76795 (p =1)	5.072 (p=1)
Single-copy	0.52465 (p = 1)	477.6 (p < 2.2e-16)
Intermediate	1.3838 (p = 1)	2.0733 (p = 1)
Multi-copy	1.8271 (p = 1)	2.1274 (p = 1)

3

THE IMPACT OF PROTEIN-PROTEIN INTERACTIONS ON THE EVOLUTIONARY AND FUNCTIONAL DIVERGENCE OF GENE DUPLICATES ANGIOSPERMS

"Not until we recognize what is holding us back, we can move forward"

3 IMPACT OF PPI ON THE DIVERGENCE OF GENE DUPLICATES

This chapter is a manuscript ready for submission to genome biology and evolution.

Jonas Defoort[1,2,3], Riet De Smet[1,2,3], Michiel Van Bel [1,2,3], Yves Van de Peer[1,2,3,4], and Lorenzo Carretero-Paulet[1,2,3]

[1]Ghent University, Department of Plant Biotechnology and Bioinformatics, 9052 Ghent, Belgium

[2]VIB Center for Plant Systems Biology, 9052 Ghent, Belgium

[3]Bioinformatics Institute Ghent, Ghent University, 9052 Ghent, Belgium

[4]Genomics Research Institute, University of Pretoria, Pretoria 0028, South Africa

3.1 ABSTRACT

Gene duplicates, either generated through WGD or SSD, are prominent in angiosperms and are believed to play an important role in generating evolutionary novelty and adaptation. Previous studies have reported differences in the evolutionary and functional fate of duplicates depending on the mechanism of duplication. For example, certain biological functions tend to be preferentially duplicated through WGD, while other functions are enriched among SSD duplicates, a pattern referred to as reciprocal retention. However, the mechanisms influencing loss and retention of gene duplicates over evolutionary time are not yet fully elucidated. Here, we investigated the impact of protein-protein interactions (PPI) in the evolutionary and functional fate of WGD and SSD duplicates in *Arabidopsis thaliana* (Arabidopsis), *Solanum lycopersicum* (tomato) and *Zea mays* (maize). Using a robust classification of gene duplicates based on phylogenetic trees and synteny analyses, a large RNAseq expression compendium, and an extensive protein interaction network from Arabidopsis, significant divergence at the level of sequence, expression pattern and protein interaction partners could be observed between tandem (SSD) and block (WGD) duplicates. Furthermore, duplicates involved in PPIs i) tend to be more evolutionary constrained in terms of expression and sequence divergence than their counterparts without interactions, and ii) are enriched in gene families predicted to be dosage balance sensitive. Our results highlight the complexity in the evolutionary dynamics and functional specialization of duplicated genes, pointing to a dominant role for the mechanism of duplication and PPIs, rather than biological functions themselves, in determining the loss and retention patterns of gene families across angiosperms.

3.2 CONTRIBUTION

- Performing the research
- Designing and performing analyses
- All Figures and tables
- Writing the manuscript
- Riet De Smet and Lorenzo Carretero-Paulet assisted in writing the manuscript
- Michiel Van Bel helped with technical analysis of genome and gene family data

3.3 INTRODUCTION

Duplicated genes are very abundant in plants and constitute a major source of evolutionary novelty and adaptation, likely playing a key role in generating phenotypic diversity and speciation [5, 222]. Duplicates can be broadly classified into two groups based on the size of the genomic region affected by the duplication. Either they are the result of a WGD, also known as polyploidizations, involving the entire genome and thus affecting all genes in the genome, or they originate form SSD, restricted to small genomic regions and mostly involving one to a few genes. Although most WGDs are followed by intense fractionation or genomic rearrangements removing from the genome most of the duplicated features, successful WGDs can be traced back at the base of the main plant lineages [222], while additional more recent events of WGD occurred independently in many lineages [21, 129, 223-225]. For example, in the widely used plant model species *Arabidopsis thaliana*, four WGD events have been detected throughout its evolution [19, 20]. The most recent ones, namely α and β events, are specific to the Brassicaceae family to which Arabidopsis belong, while the older ones, designated as γ and ε WGD events, are specific to the eudicot and angiosperm lineages, respectively. Likewise, the asterid *Solanum lycopersicum* (tomato), a model fruit-bearing crop, shares the γ and ε duplication events with Arabidopsis, and has undergone a more recent whole genome triplication estimated to have occurred 63.66 mya [21]. Finally, the monocot *Zea mays* (maize) has five detected WGD events, the shared pre-angiosperm WGD one (ε), three additional WGD events shared by all grasses, and a more recent one dated around 20.40 mya [21]. SSD events can have different origins, including tandem gene duplication, TE-based mechanisms or retroduplication, the most common one being tandem duplication originating from unequal crossing-over. Together with WGD duplicates, tandem duplicates represent the vast majority of the duplicates [5].

Previous studies have reported notable differences in the evolutionary and functional fate of duplicates depending on the duplication mode. Certain biological function categories tend to be preferentially duplicated through WGD, while other functions are enriched among SSD duplicates [29, 49]. WGD duplicates are found to be more conserved in terms of expression patterns, maintain more protein interaction partners and are retained for longer evolutionary times than SSD ones [226-228]. Furthermore, the pattern of loss and retention of duplicates within a certain gene family is consistent across angiosperms [229]. Interestingly, the role of the mechanism of duplication in the retention patterns may also be extended to other taxonomic groups such as yeasts or vertebrates [230, 231]. For example, a survey of the genomes of experimentally evolved yeast strains, reported that SSD duplicates are more functionally divergent from one another and diverge in their sub-cellular locations more than WGD duplicates [232]. Next to duplication mode, the network context has also been proposed to influence the evolutionary rate and duplicability [233, 234]. When a duplicated gene involved in network interactions is lost, it is more likely that the second member of the interacting pair gets lost too, hence restoring the copy number balance [48]. Several studies also have linked duplicate retention to the conservation of protein interaction partners [63, 232, 235] and greater difference in protein-protein interactions tends to be linked with lower expression similarity and conservation of protein domains [235].

Several theories have been proposed to explain the differential patterns of loss and retention of WGD and SSD duplicates. For example, the dosage balance hypothesis states that genomes evolve in such a way that the encoded proteins that are forming complexes or are involved in multiple steps of biological or regulatory

pathways, must remain in optimal balance [236, 237]. It is assumed that WGD duplicates do not upset stoichiometric balance of proteins in the cell because all genes in the genome are duplicated [236, 237], and are therefore preferentially retained, as their loss is expected to lead to a dosage imbalance [103]. Conversely, SSD results in one, or several, additional gene copies that are again likely to upset dosage balance and result in fitness defects, which can be resolved by either functional specialization of the duplicates, or by non-functionalization, by means of which gene copies are gradually inactivated and transformed into a pseudogene through the stochastic accumulation of mutations, being eventually removed from the genome [28]. A recently published modelling approach shows that dosage balance sensitive genes exhibit preferential duplication through WGD in angiosperms [238].

Next to dosage balance, other mechanisms for duplicate retention over longer periods of time have been put forward, including paralog interference [43], absolute dosage (higher flux) [42], dosage sub-functionalization [35], and sub-/neo-functionalization [239]. The retention forces are not indefinitely active and the function of the genes might change, leading to an escape from conservation forces and making the duplicate susceptible to loss [52, 229, 240]. It has been speculated that longer retention times set the stage for functional specialization in later time periods where they can give rise to novel functions [48, 162, 241].

However, despite intense research, the mechanisms influencing loss and retention of gene duplicates over evolutionary time are not yet fully elucidated. In this paper, we study the effect of protein-protein interactions (PPI) on duplicate gene retention and subsequent sequence and expression pattern divergence in Arabidopsis, tomato and maize. For this purpose, we generated a dataset of gene duplicates based on phylogenetic trees and synteny analyses, an extensive protein interaction network, and a large RNAseq expression compendium with uniquely mapped reads. We compare our results to two recent papers on the molecular mechanisms driving the evolutionary and functional fate of gene duplicates across 37 species of angiosperms [229, 238]. Altogether, our results support a dominant role for duplication mode and PPIs rather than biological functions themselves, in determining the loss and retention patterns of gene families across angiosperms.

3.4 RESULTS

3.4.1 Classification of gene duplicates, expression data mapping and protein-protein interactions in Arabidopsis, tomato and maize

Duplicated genes were identified using a classification of gene families across 37 angiosperm plant species [229]. Duplicates were further classified as WGD or SSD based on whether they were located in collinear regions of the genomes (block duplicates) or were found as tandems, respectively, according to i-ADHoRe [242]. The duplicates that are marked to be both tandem and block duplicates and the ones which could not be assigned to any duplication mode were grouped together and labelled 'unclassified'. This group contains a mixture of tandem, block and other duplicates and shows peaks in the Ks distribution that overlap with the WGD peaks (Figure S3.1). Phylogenetic tree based prediction shows that 66% of the unclassified Arabidopsis duplicates can be associated with WGD events (13% eudicot, 53% Brassicaceae), while only a small fraction (10%) were predicted to be recent duplications (Table S3.4) [229].

For every pair of duplicate sequences, we computed the synonymous and non-synonymous substitutions rates (Ks and Kn, respectively). As synonymous substitutions are not supposed to impact the function and/or structure of the resulting encoded protein, they stochastically accumulate throughout evolution in a neutral manner, and are thus commonly used as a proxy of evolutionary time [27]. In order to reduce the bias of synonymous substitutions saturation in old duplicates [27], only duplicates with a Ks value lower than 4 were considered for further study (Figure S3.1). In turn, the rates of non-synonymous substitutions, resulting in amino acid changes, can be used as estimates of sequence divergence (SD) between duplicates.

We used an expression dataset consisting of a compendium of RNAseq experiments comprising 56 conditions for Arabidopsis, 86 conditions for tomato, and 77 conditions for maize. Conditions include a mixture of stress conditions, tissue samples and developmental stages (Table S3.1, Table S3.2, Table S3.3). The reads were uniquely mapped and low expression filtering was applied to ensure data quality. Unlike previous studies, where mostly microarray expression data, displaying a low detection rate of duplicates, were used [226, 227], RNAseq expression data with unique mappings allow us to individually detect most of the duplicated genes in a pair (e.g. ATH1 *A. thaliana* microarrays misses probes to detect both genes in 38% of the duplicates; Table S3.5). After unique mapping of the reads, expression values were found for both duplicated genes in 63% of Arabidopsis pairs, 52% of tomato, and 48% of the maize ones. We observed significantly more block duplicates in which both genes in the pair were detected (Arabidopsis: 84%; tomato: 78%; maize 83%) than tandem duplicates (Arabidopsis: 33%; tomato 32%; maize 27%) (hypergeometric test p value; Arabidopsis: $p < 2.2e\text{-}16$; tomato: $p < 2.2e\text{-}16$; maize: $p < 2.2e\text{-}16$). This is likely due to the large number of young tandem duplicates without unique mapping caused by the little or no sequence divergence among them (Figure S3.2)

We assembled a compendium of experimental Arabidopsis PPIs based on small- and large-scale experiments. The compendium consists out of 52,613 interactions for 10,266 proteins. In the Arabidopsis duplicate set, 136 duplicates interact with each other, in 1186 duplicates both proteins have interaction partners, in 1581 only one of the duplicates has interaction partners and 2329 duplicates are without interaction partners. The set without interactions still contains a high number of false negatives due to the lack of experimental studies. To investigate the influence of PPI on tomato and maize duplicates, we projected the experimental

Arabidopsis PPIs onto their corresponding orthologous genes. If an Arabidopsis gene has at least one PPI, all tomato and maize genes within the same gene family are assigned to the category with PPI. For tomato, this results in 2492 duplicate pairs with and 3182 duplicate pairs without PPIs and for maize there are 3670 duplicate pairs with and 6984 duplicate pairs without PPI.

3.4.2 WGD duplicates show stronger conservation in terms of expression and interaction partners than SSD duplicates

To assess differences in functional conservation between duplicates, we used expression divergence (ED) and protein interaction divergence (ID) as proxies [235]. For each of the duplicate pairs, the ED was calculated as the relative number of conditions in which only one of the duplicates is detected. ID, in turn, was calculated as 1 minus the retention rate, defined as the number of interaction partners that are shared between two duplicates divided by the sum of unique interaction partners of both duplicates. In order to reduce the noise due to the fact that not all proteins have been experimentally investigated, ID was only calculated for duplicates in which one of the duplicates has at least four PPI and the other duplicate at least one PPI. 788 pairs were found to be above this cut-off.

Arabidopsis, tomato and maize block duplicates show significant lower ED than tandem duplicates (Figure 3.1), which is in agreement with previous observations in Arabidopsis and rice [226, 227]. Similarly, ID rates are lower among block duplicates than among tandem ones (Figure 3.2A). There are more block duplicates (23%) with more than half of the interaction partners conserved, compared to only 6% for tandem duplicates (Fisher exact test; p=1.2e-8). We also found more tandem duplicates without any shared interaction partners (48%) than block duplicates (30%) (Fisher exact test; p=2.3e-2). Finally, although ID likely represents an underestimation of the number of shared interaction partners, our analyses were replicated using different cut-offs (from at least one up to 14 interaction partners in one the duplicates), always resulting in significant differences between tandem and block duplicates (Figure S3.3).

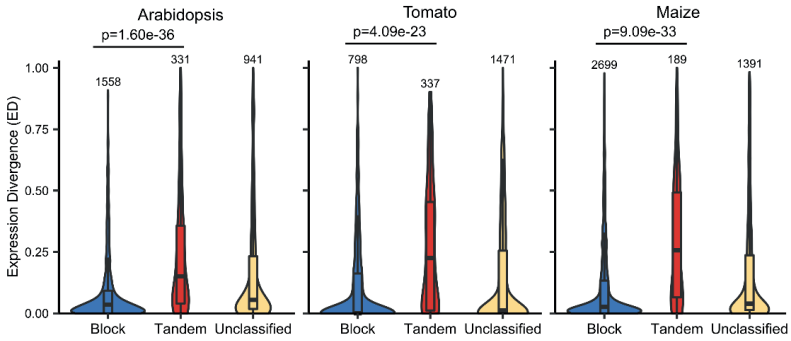


Figure 3.1: ***Expression divergence (ED) between Arabidopsis, tomato and maize duplicates per duplication mode.*** *Violin plots and embedded boxplots for each duplication mode and species are shown. P-values resulting from Wilcoxon's rank sum tests of the differences between block and tandem duplicates are shown. Number of duplicates are shown above each plot.*

To get insights into the evolution of ED and ID over time, we plotted both estimates for Arabidopsis pairs of duplicated genes as a function of Ks, used here as a proxy of evolutionary time since duplication. With increasing Ks, both ID (Figure 3.2B) and ED (Figure 3.2C) increase faster in tandem than in block duplicates.

Statistical analysis revealed a significant correlation between ED and ID (Pearson correlation test; correlation: 0.13, p value: 6.49e-5), which hints that duplicates with larger expression divergence tend to lose their interaction partners. Tandem duplicates show stronger correlation than block duplicates (Figure S3.4; Pearson correlation test; correlation block: 0.13, p value: 1.1e-3; correlation tandem: 0.27, p value: 3.8e-2). These results, taken together, strongly suggest that tandem duplicates tend to be more functionally diverged both in expression and in interaction partners than block duplicates.

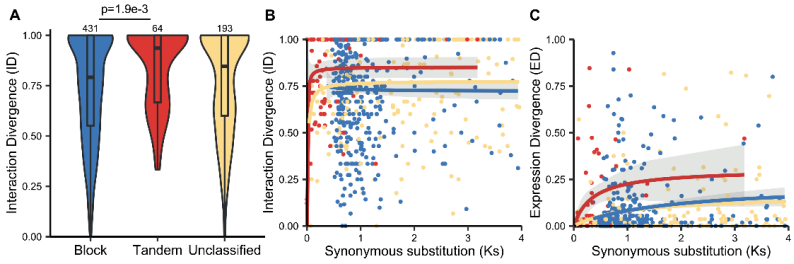


Figure 3.2: ***Evolution of interaction and expression divergence in Arabidopsis.*** *A) Violin plots of ID between duplicates per duplication mode. The corresponding boxplots are embedded. The number of duplicates is shown on top of each boxplot. ID was only calculated for duplicates with both interactions and one copy having at least 4 interaction partners. B) ID plotted as a function of Ks. C) ED plotted as a function of Ks. In order to reduce the effect of nonsynonymous substitution saturation, a Michaelis-Menten-type saturation curve was fit to each group independently with 95% confidence regions indicated as grey areas.*

3.4.3 PPIs constrain expression and sequence divergence of tandem and block duplicates

To further explore the influence of PPIs on sequence and expression divergence of pairs of duplicated genes, we compared duplicates with and without PPI in all three species. PPIs were found to be overrepresented among block duplicate genes (Arabidopsis p < 2.2e-16, Tomato p < 2.2e-16; maize p= p < 2.2e-16), while underrepresented among tandem duplicate ones (Arabidopsis p=5.25e-11, Tomato p=6.68e-8; maize p= p < 2.2e-16) (Fisher exact test on count data with Benjamini-Hochberg multiple testing correction). Furthermore, Arabidopsis duplicates without PPIs exhibit significantly higher ED than duplicates with interactions in both block and tandem duplicates (Figure 3.3A, Wilcoxon rank sum test; block p=4.60e-13, tandem p=1.63e-3), although there is also a significant difference between the duplication modes (Wilcoxon rank sum test; p < 2.2e-16). After projecting Arabidopsis PPIs data onto the corresponding orthologs in tomato and maize proteomes, we detect the same significant results for duplicates with and without PPI, both within and between duplication modes (Figure 3.3B&C, Wilcoxon rank sum test, Tomato: block p=7.91e-5, tandem p=5.6e-3, block vs. tandem p < 2.2e-16; Maize: block p < 2.2e-16, tandem p=0.92e-3, block vs. tandem p < 2.2e-16). To get additional insights into this observation, we further split Arabidopsis duplicates into four categories; self interacting duplicates, i.e., homodimers, duplicates in which both genes have detected PPI, duplicates in which only one copy has detected PPIs and duplicates without any detected PPI. Self-interacting duplicates showed the lowest ED followed by duplicates in which both copies showed interaction partners (Figure S3.5). Pairs in which only one of the duplicates bears interactions resulted in significantly higher ED than pairs in which both copies showed interaction partners. Finally, duplicates without PPI showed significantly higher ED than all other three categories.

Next, we examine the GO functional terms annotating Arabidopsis genes belonging to each of the duplication modes partitioned by those showing PPIs or not. A reciprocal pattern of enrichment in GO molecular functions could be observed (Figure 3.3D). Block duplicates with PPI are enriched for binding (protein, DNA and RNA), enzymatic activity (kinase, transferase, catalytic) and signal transducing activity. Tandem duplicates with PPI are also enriched for carbohydrate binding, which is in turn underrepresented in block duplicates with PPI. Block duplicates without PPI are enriched for catalytic activity and hydrolase, which are also overrepresented among tandems without PPI. Tandem duplicates are enriched in transporter activity. Both tandem and block duplicates without PPIs are overrepresented in categories linked to enzymatic activity. The GO enrichment analysis for tomato and maize duplicates showed similar results, likely because PPI data were originally obtained from their corresponding Arabidopsis orthologs (data not shown).

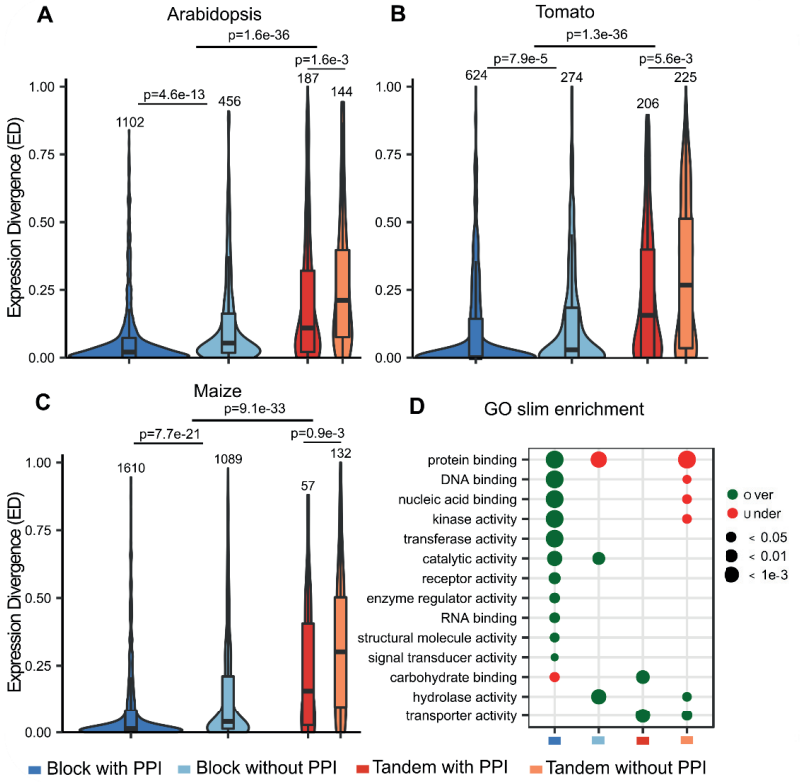


Figure 3.3: ***Expression divergence and GO functional enrichment analysis of duplicates with and without PPI.*** *Violin plots of expression divergence for Arabidopsis A), tomato B) and maize C) duplicates with and without PPIs. The corresponding boxplots are embedded. D) Enrichmemt analysis of GO molecular functions belonging to the plant GO slim category for Arabidopsis block and tandem duplicates with and without PPI. Only experimentally validated GO annotations were considered. GO terms significantly under- and over-represented (p-value < 0.05 hypergeometric test + BH multiple testing correction) are plotted.*

In order to further examine the impact of PPIs in differences of ED and SD between duplicates over evolutionary time, the later expressed as estimates of non-synonymous substitution rates (Kn), we plotted ES and SD as a function of Ks. In all three species, we observe larger ED both for tandem and block duplicates without PPI with respect to those with PPIs (Figure 3.4). However, results were less clear in tomato and maize, likely due to the sparser distribution of duplicates with Ks higher than 1. This, together with the high error

rate of false negatives and positives to be expected from the projection of PPI data from Arabidopsis, make more difficult to unambiguously interpret the results. However, when we plotted the average ED per Ks bins (Figure S3.6), differences we observe even clearer differences between duplicates with and without PPI both in tandem and block duplicates. In terms of SD, we observed higher Kn in duplicates without PPI compared to those with PPI, for both duplication modes and for all three species (Figure 3.4). The observations in Arabidopsis for SD and ED, together with the similar trends found in tomato and maize, highlight the slower divergence rates of duplicates with PPI compared to the ones without PPI, a trend that can be observed both between block and tandem duplicates.

Finally, in order to assess whether the actual number of PPI partners of duplicates was linked to their divergence in terms of expression or sequence, we plotted ED and SD of Arabidopsis *versus* the unique number of interaction partners for both modes of duplication combined (Figure S3.7). A small, although significant, downward trend in ED and SD in parallel to the increase in the number of interaction partners could be observed (ED; Spearman correlation = -0.078, p value 2.9e-5) (SD; Spearman correlation = -0.084, p=2.267e-8). This suggests that the actual number of PPIs has only a minor, although significant, influence on the ED and SD between duplicates.

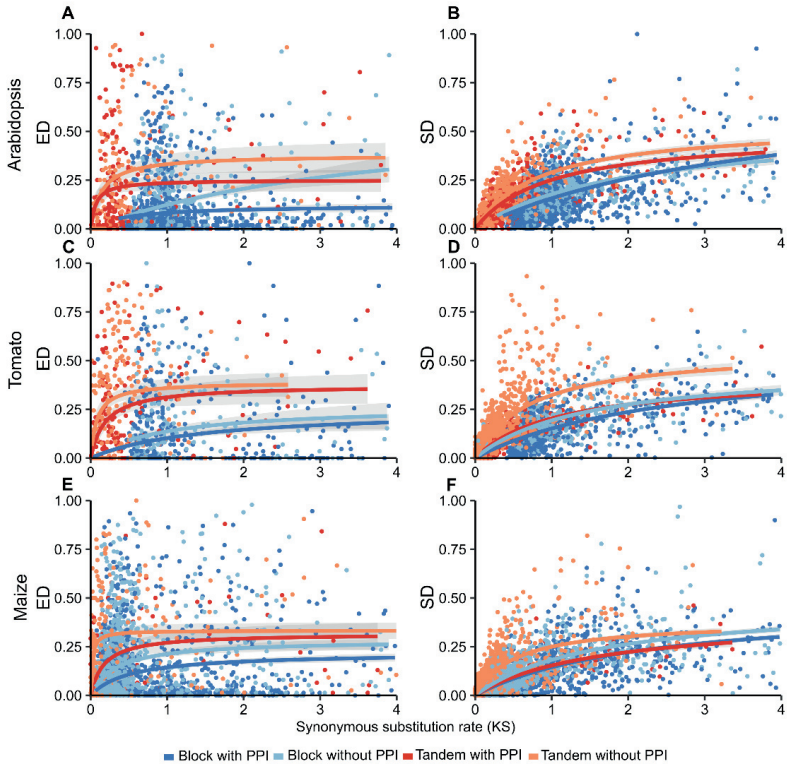


Figure 3.4: ***Evolution of sequence and expression divergence of block and tandem duplicates with and without PPI.*** *ED and SD are plotted as a function of Ks. In order to reduce the effect of nonsynonymous substitution saturation, a Michaelis-Menten-type saturation curve was fit to each group independently with 95% confidence regions indicated as grey areas.*

3.4.4 PPI and duplication mode may help to explain the duplicate retention patterns observed across angiosperms

A recent study revealed strikingly similar retention and loss patterns in gene families across 37 angiosperm species with fully sequenced genomes [229]. Genes within a certain gene family seem to have a preferential copy number to which they return if sufficient time passes. Using a clustering approach, gene families including representatives from at least 32 out of the 37 species (i.e., core gene families), were further subdivided into three groups according to their retention patterns: “single copy” gene families, which return “quickly” after duplication back to the single copy status; “intermediate” gene families, in which duplicated genes are maintained for longer periods, but eventually get lost; and “multi-copy” gene families, in which duplicates are maintained for much longer periods. Gene duplication modes are unequally distributed among groups. Block duplicates are mostly found within the intermediate and multi-copy gene families. 41% and 34% of all Arabidopsis block duplicates, 35% and 39% of all tomato block duplicates, and 37% and 26% of all maize block duplicates, respectively, were found among those groups of gene families. In contrast, tandem duplicates are mostly found within the multi-copy and non-core gene families. 20% and 62% of all Arabidopsis tandem duplicates, 18% and 68% of all tomato tandem duplicates, and 8% and 72% of all maize tandem duplicates, respectively, were found among those groups of gene families. Comparison of the duplication mode of Arabidopsis duplicates to the duplication mode of genes in other angiosperms belonging to the same gene family reveals that block duplicates from Arabidopsis, tomato and maize also duplicate mostly through WGD in other species, while tandem duplicates also duplicate mostly through tandem duplication (Figure 3.5). Arabidopsis block and tandem duplicates with PPI show a higher percentage of block duplicates compared to the duplicates without PPI in the other species (Figure 3.5), suggesting a role for PPI. The percentages show that no straight cut can be made between gene families that either duplicate through tandem or WGD duplication. Tandem duplicates do appear in the preferential WGD families and vice versa.

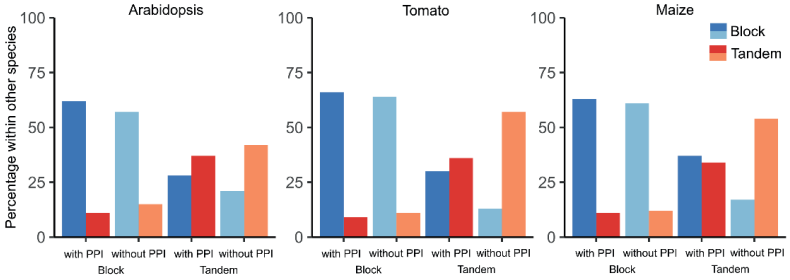


Figure 3.5: ***Duplication modes are conserved across angiosperm gene families:*** *The bars represent the total percentage of Arabidopsis, tomato and maize duplicates, partitioned by mode of duplication and occurrence of PPIs, which duplicate through tandem or WGD within the same gene family in other 36 other angiosperm species.*

We next used this classification of gene families in order to assess whether the effect of PPIs on the loss and retention patterns of duplicates observed in Arabidopsis could also be extended to other angiosperms. We observe that PPI are overrepresented among duplicates belonging to intermediate (Arabidopsis p=4.06e-15; Tomato p < 2.2e-16; maize p < 2.2e-16) and multi-copy groups (Arabidopsis p=4.53e-16; Tomato p < 2.2e-16; maize p < 2.2e-16) (Fisher exact test + BH correction), in agreement with previous observations [229]. From single to multi-copy group, i.e., with increasing duplicate retention, ID and ED decrease (Figure S3.8). When

we examine the evolution of ED *versus* Ks in the different retention groups, we observe a general trend among both tandem and block duplicates without PPI to diverge faster than the ones with PPI (Figure 3.6). Only the block duplicates in the intermediate group seem to diverge from this pattern. The difference between tandem and block duplicates is also observed in each group (Figure 3.6).

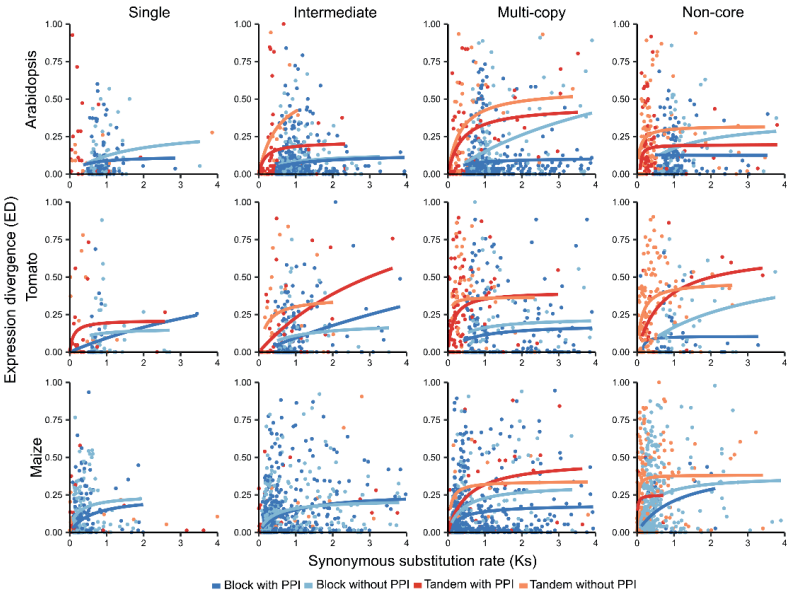


Figure 3.6: ***Evolution of expression divergence per retention group****. Gene families were classified according to their respective retention patterns as single, intermediate or multi-copy as reported in Li, 2016. A fourth group of gene families (i.e., non-core) comprised gene families without representatives in at least 32 out of the 37 species examined. ED of duplicates within gene families belonging to each retention group is plotted as a function of Ks. In order to reduce the effect of nonsynonymous substitution saturation, a Michaelis-Menten-type saturation curve was fit to each group independently with 95% confidence regions indicated as grey areas. In some groups no function could be fit to tandem duplicates due to lack of data points.*

3.4.5 Duplicates with PPI are enriched among reciprocally retained angiosperm duplicates

In order to further substantiate these observations, we compared our results to the ones recently reported in [238], where a modelling approach was used to rank angiosperm genes families resulting from the 37 species classification based on their preferential retention after WGD. This pattern, referred to as reciprocal retention, is hypothesized to stem from constraints on the dosage balance of the genes concerned with their interaction context. According to this rank, gene families were classified as top or bottom depending on whether they showed preferential retention after WGD or not, respectively. We examined these families for their modes of duplication, PPIs, ED and SD. As expected, Arabidopsis block duplicates are predominantly found among top gene families, while tandem duplicates are mostly found among bottom gene families (Figure 3.7). Duplicates within top ranked genes families are enriched for PPIs (hypergeometric test; p=1.78e-15). Both tandem (p=1.23e-06) and block (p=4.46e-02) duplicates belonging to gene families in the top ranked group are enriched for PPI. Likewise, the number of duplicates with PPI gradually decreases from the top to the bottom gene families (Figure 3.7). This suggests that the influence of PPIs in the reciprocal

retention patterns of duplicates is similar for both duplication modes. Moreover, the tandem duplicates in the top ranked gene families, so gene families which preferentially duplicate trough WGD, show significant less expression divergence than tandem duplicates in the bottom categories which preferentially duplicate though SSD (Wilcoxon rank sum test, Arabidopsis p=1.325e-06; tomato p=7.596e-07; maize p=7.93e-3) (Figure 3.8). This suggests that the ED of tandem duplicates is dependent on the gene family to which they belong.

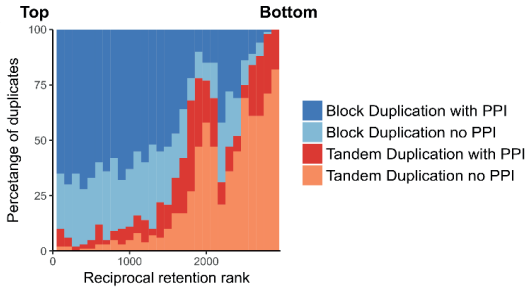


Figure 3.7: ***Distribution of duplication modes, with and without PPI, and reciprocal retention rank*** *[238]. The stacked histogram shows the percentage of duplicates from each category plotted as a function of the reciprocal retention rank.*

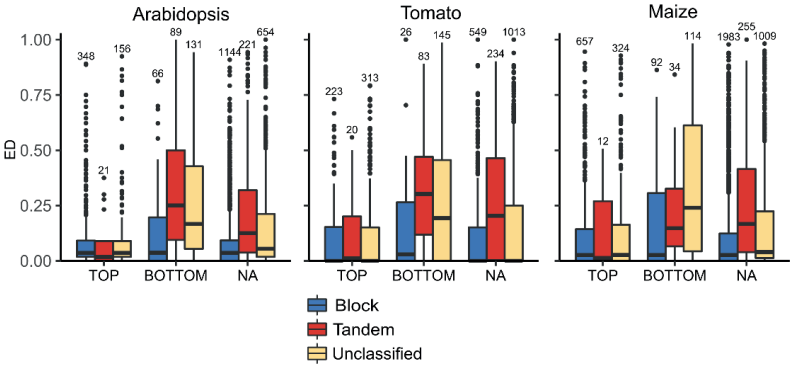


Figure 3.8: ***Expression and sequence divergence of duplicates belonging to top and bottom gene families partitioned by duplication mode.*** *Box plots of ED (A) and SD (B) between duplicates, with the corresponding boxplots embedded. The number of duplicates is shown on top of each plot.*

3.5 DISCUSSION

Despite intense research, the molecular mechanisms underlying the evolutionary and functional fate of genes after duplications are not yet fully elucidated. Specific properties of genes and, in a broader context, of gene families, have already been reported to influence the divergence patterns of duplicates, prominently the duplication mode, the biological function encoded, as well as the species [29, 49, 228, 229, 232]. In this paper, we studied the interplay between PPIs and the mechanism of duplication on subsequent sequence and expression pattern divergence in three angiosperm plants with different histories of SSD and WGD. By using a uniquely mapped RNAseq compendium from multiple species with a variety of samples, we were able to detect the majority of the duplicates in a more robust and reliable way compared to previous studies using microarray data (Table S3.5) [226, 227], although there is still some room for improvement to detect duplicates in the lower Ks regions (Figure S3.2). Furthermore, we tried to overcome the lack of experimental

PPI data in plants by projecting the Arabidopsis PPI network onto the corresponding orthologs in tomato and maize. Because of this transfer through orthology, it is impossible to estimate the ID in tomato and maize and also the resolution in duplicates with a higher Ks values and tandem duplicates is lower.

Despite these limitations, our analysis revealed similar evolutionary dynamics for Arabidopsis, tomato and maize duplicates. Tandem duplicates tend to diverge faster than block duplicates at the sequence, expression and shared PPI divergence levels, which may be linked with their timeframe of contribution to adaptation. It is generally accepted that the cost associated with the maintenance of additional gene copies results in the loss of most duplicated genes by means of non-functionalization or pseudogenization, while only a minor fraction of duplicates expected to be maintained through the acquisition of novel or specialized functions [222]. The faster evolutionary rates observed for tandem duplicates limits their potential in evolution, which might explain why tandem duplications are often found in categories related to stress where they can have an immediate effect [243]. In contrast, block duplicates will be diverging at slower rates, which leads to longer retention times and increased potential for evolutionary innovation and adaptation [162, 244]. This might explain why successful WGD events can be often found associated with periods of increased environmental stress and/or fluctuations, such as the one around the Cretaceous-Paleogene (K-Pg) extinction event about 66 million years ago [21, 23, 24, 107]. Furthermore, allopolyploidy (polyploidization resulting from interspecies hybridization of the same genome) events in grasses seem to have led to the dominance of C4 grasses over C3 ones and their subsequent territorial expansion [245]. However, the contribution of tandem duplication might have been underestimated due to miss-annotation/assembly of tandems as a single gene [246], a problem that would be specially affecting genomes of draft/poor quality (Panchy, et al. 2016).

A second characteristic that it is accepted to have an important impact in the evolution of duplicates is the encoded biological function. The link between duplicability and functional categories has been described in multiple studies [29, 49, 228, 229, 247]. Some studies claim a link between duplication mode and functional retention, while others rather suggest their independence. Transcription factors, genes involve in signalling and interacting genes were found to be enriched among retained duplicates and genes involved in maintenance of DNA repair and integrity, and organellar function were found to lose the duplicate copies [229]. A similar retention pattern has been observed across 17 fungi genomes [230] and also large-scale analysis of prokaryotic genomes revealed that the same functional categories seemed to be linked with the size of the genome [149, 150]. This suggests that the function-biased retention of duplicated genes might be present across all species. This pattern, referred to as reciprocal retention, has been linked to the differential dosage balance sensitivity of WGD and SSD duplicates (Tasdighian, et al. 2017). It remains to be established whether gene function directly influences gene duplicability or whether biased gene retention is the by-product of other evolutionary phenomena instead. These phenomena could be the molecular and/or network properties of proteins which are plaining an important role in retention according to the theories such as dosage balance.

Our results also support PPI as a third characteristic of duplicated genes that may result in longer retention times independently of the duplication mode. First, duplicates with PPIs show a lower SD compared to those without PPI. Differences in non-synonymous substitutions are, however, small, since only a small fraction of the amino acids (mostly at the surface) are supposed to have an effect on interaction and structural properties of proteins [248]. Second, PPIs were found to be overrepresented among block duplicate genes

and in gene families that preferentially duplicate though WGD. Similarly, PPI were found to be overrepresented among genes retained for longer times (intermediate and multi-copy group) in a thorough classification of gene families across 37 angiosperm genomes [229]. In general, duplicates with PPI show lower expression divergence than those without PPI. Only in the intermediate group a different pattern is observed (Figure 3.6). This group was suggested to be dosage balance sensitive and maintains the duplicate genes for a longer time, but eventually most of them get lost [229]. The curve of the intermediate duplicates without PPI follows the curve of those with PPI almost exactly. This could be explained by either the duplicates without PPI are in the wrong category due to the lack of study or other mechanisms are acting upon them. Third, PPI may be constraining the divergence of both tandem and block duplicates, an observation that may be linked to the gene family to which they belong to. Tandem duplicates in a gene family that preferentially duplicate through WGD were found to be less diverged then tandem duplicates found within families that preferentially duplicate through SSD (Figure 3.8). Based on their reciprocal retention pattern these duplicates are thought to be dosage sensitive [238]. Potentially, this could be the result of a loss of the WGD duplicate in a dosage sensitive gene family, which is buffered by a tandem duplication of that gene.

The GO enrichment analysis revealed that categories that are often associated with dosage balance sensitivity (binding of proteins, DNA and RNA) [110] are overrepresented among block duplicates with PPI and underrepresented among duplicates without PPI. These categories had been previously shown to be retained after WGD [29, 49, 229]. Both tandem duplicates (with and without PPI) and block duplicates without PPI are overrepresented in categories linked to enzymatic activity and transport activity. The retention of these groups can be linked to absolute dosage constraints [42, 52]. In contrast to the relative dosage as stated by the dosage balance hypothesis, duplication retention may also result from selection on the absolute dosage of certain gene products, i.e., the concentration of a protein in a cell. A higher concentration could generate a higher throughput of the corresponding pathway, referred to as metabolic flux. For most pathways in which a single enzyme increase has no influence on the flux, WGD could provide an increase by duplicating all components (e.g. catalytic activity) [52]. Enzymes that are working independently or that provide a bottleneck in the pathway could take advantage of a SSD (e.g. hexose transport in yeast [249]) [250].

Despite their independent histories of WGD and SSD, the similarities found in terms of gene loss and retention patterns across 37 angiosperm species [229], suggest that the evolutionary forces guiding duplicate divergence are widespread. In summary, we report here, that, i) PPIs constrain sequence and expression divergence of duplicates in both tandem and block duplicates in Arabidopsis, tomato, and maize. ii) PPIs influence the loss and retention patterns observed across all angiosperm and iii) the reciprocal retention pattern of duplicates with PPI shows signs of dosage sensitivity. In addition, we provide further support for the difference in expression and protein interaction partner divergence between block and tandem duplicates and the correlation between expression/sequence divergence and interaction partner loss. Altogether, our results highlight the complexity in the evolutionary dynamics and functional specialization of duplicated genes, pointing to a dominant role for the mechanism of duplication and PPIs (considered as properties of gene families), rather than biological functions themselves, in determining the loss and retention patterns of gene families across angiosperms.

3.6 MATERIALS AND METHODS

3.6.1 Classification of block and tandem duplicates

Block duplicates, putatively arising from WGD events, and tandem duplicates, conforming the majority of SSD duplicates, were detected using the PLAZA framework [251]. A new PLAZA instance was built for the gene families reported in [229] using the default workflow as described for PLAZA 3.0 [242]: 1) gene homology was determined by applying MCL clustering to the all-vs-all BLASTP results; 2) tandem duplicates were defined as any homologous genes located within a 30 genes window distance of the same chromosome; 3) block duplicates were detected on the basis of their syntenic arrangements in the genome by means of the IADHoRe program [252].

3.6.2 Estimates of synonymous and non-synonymous substitution rates

Estimates of synonymous substitution rate (Ks) and non-synonymous substitution rate values (Kn) were obtained for each pair of paralogous genes on the basis of the ClustalW alignment of the protein coding sequences [192] using parameter recommendations from [253]. PAL2NAL [254] was used to back-translate the aligned amino acids into the corresponding codons without gaps. Then, codeml [180] from PAML [255] was used to obtain Ks and Kn values for each gene pair using the GY model [180], with stationary codon frequencies empirically estimated by the F3x4 model. Within this study we use the Kn value as a sequence divergence (SD). Estimates of values were obtained for all paralogous pairs associated with the predicted duplication events inferred by the gene tree/species tree reconciliation process. For cases where there are multiple possible pairs for a predicted duplication event, we calculated KS-values for all possible gene pairs and selected the gene pair with the smallest KS-value to represent the timing of the duplication event [229].

3.6.3 RNAseq compendium and expression measures

The Arabidopsis RNAseq expression compendium consists out of 56 conditions downloaded from Cornet 3.0 (Table S3.1) [256]. In contrast, the tomato and the maize RNAseq expression compendia were taken from the NCBI's Sequence read archive, and consists out of 84 and 77 conditions, respectively (Table S3.2 & Table S3.3). All three datasets were analysed using the same pipeline. Trimmomatic [257] was used to perform quality filtering and adaptor removal. The reads were mapped using GSNAP [258], only the uniquely mapped reads where retained. Gene counting was done with Htseq-count [259] after which a transformation to CPM was done with EdgeR [260]. To ensure data quality, low expression filtering was performed. Genes with a sum expression count over all conditions lower than 2 times the number of conditions were removed. In total, 19318 Arabidopsis, 19495 tomato, and 23164 maize genes were mapped. 63% of the Arabidopsis, 52% of the tomato, and 48% of the maize duplicates have expression detected for both duplicates. The expression divergence (ED) was defined as the relative number of conditions in which only one of the duplicates is detected (C_1 and C_2), divided by the total number of conditions in which they are detected (C).

$$ED = \frac{C_1 + C_2}{C}$$

This measure takes into account the number of conditions in which the duplicates are expressed and reduces differences due to the combination of the different experiments. A measure of 0 means that both duplicates

are always expressed in the same conditions. A measure of one means that the duplicates were never detected together.

3.6.4 Protein-protein interaction network and measures

A compendium of protein-protein interactions in Arabidopsis was constructed combining the following sources: BioGRID 3.4 [261], Arabidopsis Interactome [63], MIND [262] CORNET [263] (only experimentally validated interactions), STRINGv9.1 (only category Binding) [62], EVEX [264] (only category binding) and a TAP data set assembled from literature [205-208, 210-212, 214, 215, 217-219, 265, 266]. After removing redundancy and self-interactions, we obtained a set of 52613 interactions for 10266 proteins. The Arabidopsis PPIs were transferred to the corresponding orthologous genes in tomato and maize based on gene family membership. If at least one interaction was present in one of the Arabidopsis genes, all tomato and maize genes were assigned to the category with PPI. The Interaction divergence (ID) between 2 duplicates is calculated as one minus the retention rate, which in turn is defined as two times the number of interaction partners shared between two duplicates ($I_{1,2}$) divided by the sum of interactions in each of the duplicates (I_1, I_2).

$$ID = 1 - \frac{2I_{1,2}}{I_1 + I_2}$$

The ID was only calculated for duplicates with at least one PPI partner in a copy and at least four in the second copy (Figure S3.3).

3.6.5 Computational resources

All data processing and data analysis was done using a combination of Python and R scripts. The read mapping and counting pipeline was run for all samples in parallel on a computing cluster with Linux computing nodes (2.4GHz, Intel). All samples were run individually on 4 nodes with each 4GB (Arabidopsis and tomato) or 6GB (Maize) of memory. The running time per sample ranged between 1 and 8 hours depending on the library size.

3.7 SUPPLEMENTAL DATA

3.7.1 Supplemental figures

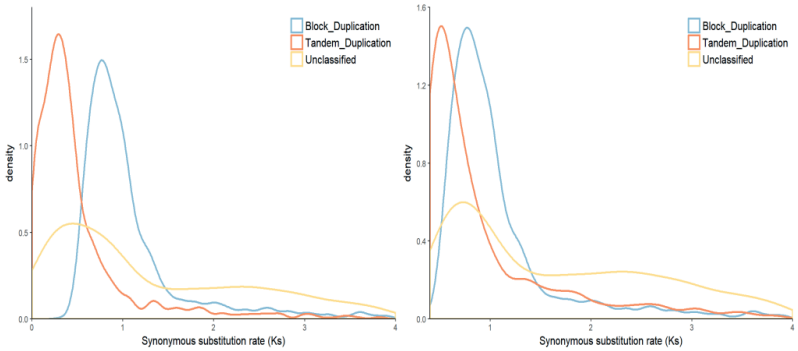
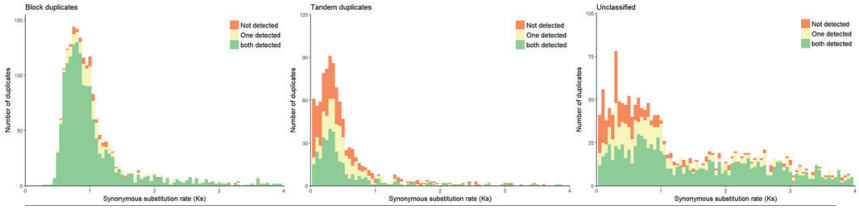


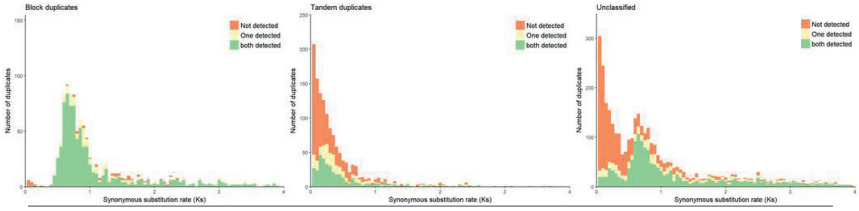
Figure S3.1: synonymous substitution rate (Ks) distribution of the duplicates for each Duplication mode in A. thaliana. Left: full set of duplicates. Right: Filtered set with a Ks value higher than 0.1 and lower than 4 values in order to reduce the bias towards young/non-diverged duplicates and the influence of synonymous substitutions saturation in old duplicates.

Arabidopsis

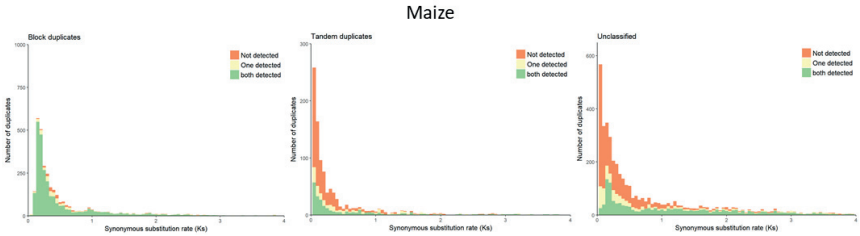


A. thaliana	Not detected	One detected	Both detected	Total	% Both detected
Block duplicates	77	230	1612	1919	84%
Tandem duplications	486	274	370	1130	33%
Unclassified	379	508	1296	2183	59%
Total	**942**	**1012**	**3278**	**5232**	**63%**

Tomato

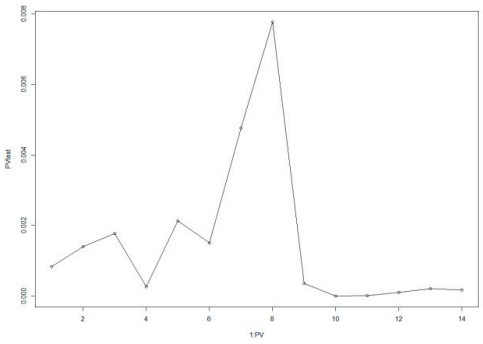


S. lycopersicum	Not detected	One detected	Both detected	Total	% Both detected
Block duplicates	75	237	1074	1386	78%
Tandem duplications	779	340	529	1648	32%
Unclassified	943	385	1312	2640	50%
Total	**1797**	**962**	**2915**	**5674**	**52%**



Z. mays	Not detected	One detected	Both detected	Total	% Both detected
Block duplicates	237	337	2826	3400	83%
Tandem duplications	1024	218	450	1692	27%
Unclassified	2811	933	1818	5562	33%
Total	**4072**	**1488**	**5094**	**10654**	**48%**

Figure S3.2: Detection of the Duplication modes in function of the Ks value within Arabidopsis, tomato and maize. Below the plots a table with the exact numbers is show.



IRR cut-off	# block	# tandem	PVtest
1	657	49	8.34E-04
2	577	44	1.40E-03
3	501	38	1.77E-03
4	430	32	2.63E-04
5	386	27	2.13E-03
6	337	22	1.51E-03
7	312	20	4.76E-03
8	286	19	7.76E-03
9	261	18	3.55E-04
10	234	16	3.65E-06
11	223	15	1.55E-05
12	211	14	1.00E-04
13	196	14	2.09E-04
14	177	14	1.75E-04

Figure S3.3: Significance of the interaction divergence rate comparison between tandem and block duplicates for increasing number of PPI cut-off (x-axis). Table shows the number of block and tandem duplicates with are retained after the cut-off and the significance of the comparison. This shows that the comparison is significant independent of the cut-off. We selected min 4 PPI cut-off based on the first valley.

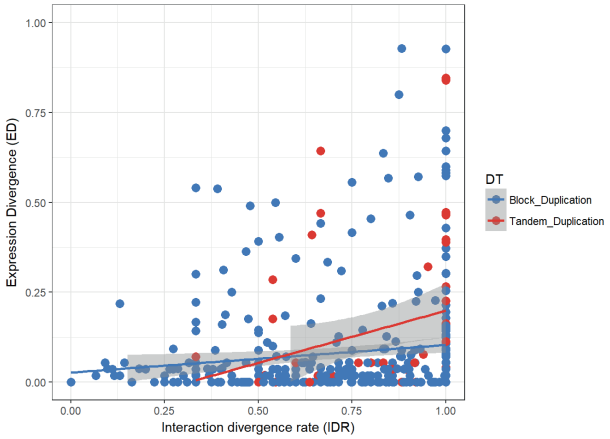


Figure S3.4: Expression divergence (ED) versus interaction divergence (ID) in Arabidopsis split up between block duplicates and tandem duplicates.

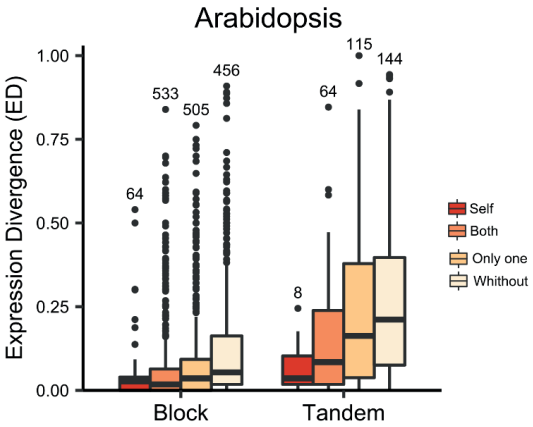


Figure S3.5: ***Expression divergence for Arabidopsis duplicates.*** *Self: self-interacting duplicates; Both: both duplicates have detected PPI; One: only one duplicate has detected PPI; Without: none of the duplicates has detected PPI. The number of duplicates is depicted above the boxplot.*

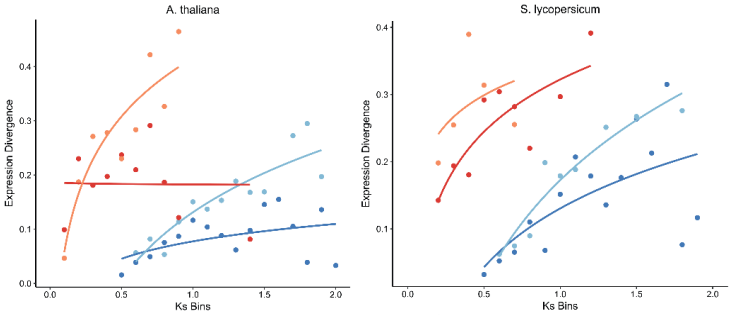


Figure S3.6: Expression divergence versus synonymous substitution rate bins. Each dot represents the average expression divergence of all duplicates within that bin (Binsize = 0.1)

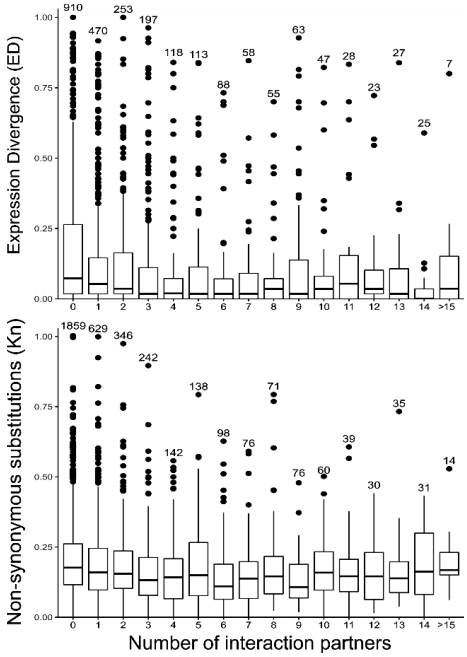


Figure S3.7: Boxplots of expression divergence (ED) and non-synonymous substitutions rate (Kn) per number of unique protein interaction partners for both duplicates in Arabidopsis.

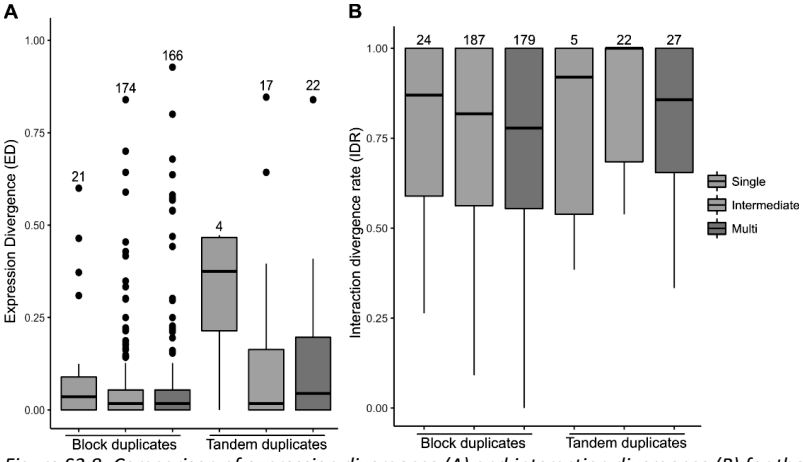


Figure S3.8: Comparison of expression divergence (A) and interaction divergence (B) for the different retention groups. First three are block duplicates, next three are tandem duplicates.

3.7.2 Supplemental tables

Table S3.1: List with RNAseq experiments in the expression compendia for A. thaliana.

Table S3.2: List with RNAseq experiments in the expression compendia for Tomato.

Table S3.3: List with RNAseq experiments in the expression compendia for Maize.

Table S3.4: Number of duplicates in each Duplication mode. Eudicot/Brassicaceae/Recent= age category based upon tree prediction.

Table S3.5: Comparison of duplicate detection in A. thaliana between ATH1 microarray and RNAseq compendium. First table represents the number and percentage of duplicates detected in the RNAseq compendium after filtering. Second table represent the presence of probes for the duplicates on the ATH1 microarray. The third table shows the comparison between the detected set and the presence on the microarray. This show that 17% of the duplicates are undetectable and for 21% of our duplicates only one gene can be detected using microarrays.

4

THE ‘TRANSEQ’ 3’ END SEQUENCING METHOD FOR HIGH-THROUGHPUT TRANSCRIPTOMICS AND GENE SPACE REFINEMENT IN PLANT GENOMES

`Less is more`

4 THE 'TRANSEQ' 3' END SEQUENCING METHOD

This chapter is a technical advances article resubmitted to 'the plant journal'.

Oren Tzfadia*[1,2,3], Samuel Bocobza*[4], **Jonas Defoort[1,2,3]**, Efrat Almekias-Siegl4, Matan Levy[4], Veronique Storme[1,2,3], Stephane Rombauts[1,2,3], Diego Adhemar Jaitin[5], Hadas Keren-Shaul[5], Yves Van de Peer[1,2,3,6], Asaph Aharoni[4].

[1]Center for Plant Systems Biology, VIB, Ghent, Belgium

[2]Department of Plant Biotechnology and Bioinformatics, Ghent University, Ghent, Belgium

[3]Bioinformatics Institute Ghent, Ghent University, Ghent, Belgium

[4]Department of Plant Sciences, Weizmann Institute of Science, Rehovot, Israel

[5]Department of Immunology, Weizmann Institute of Science, Rehovot, Israel

[6]Genomics Research Institute (GRI), University of Pretoria, 0028 Pretoria, South Africa

*These authors contributed equally to this work

4.1 ABSTRACT

High-throughput RNA sequencing has proven invaluable not only to explore gene expression, but also for both gene prediction and genome annotation. However, RNA sequencing, carried out on tens or even hundreds of samples, requires easy and cost-effective sample preparation methods using minute RNA amounts. Here, we present TranSeq, a high-throughput 3'-end sequencing procedure that requires 10- to 20-fold fewer sequence reads than the current transcriptomics procedures. TranSeq significantly reduces costs and allows a great increase in size of sample sets analysed in a single experiment. Moreover, in comparison to other 3' end sequencing methods reported to date, we demonstrate the reliability and immediate applicability of TranSeq and show that it not only provides accurate transcriptome profiles but produces the potential to detect expression of specific gene family members possessing high sequence similarity. Furthermore, mapping TranSeq reads to the reference tomato genome facilitated the annotation of new transcripts improving > 45% of the existing gene models. Hence, TranSeq is anticipated to boost large-scale transcriptome assays and impact the spatial and temporal resolution of gene expression data and their visualization, in both model and non-model plant species. Moreover, as already done for tomato (ITAG3.0; www.solgenomics.net), we strongly advocate its integration into current and future genome annotations.

4.2 CONTRIBUTION

- Analysis of expression data and gene family data
- Comparison between TruSeq and TranSeq for duplicate detection
- Figures: 4.4 & 4.5
- Supplementary Figures: S4.4 & S4.5
- Supplementary Tables: S4.3
- Assisting in writing the manuscript

4.3 INTRODUCTION

The revolution in whole genome and transcriptome sequencing added a new dimension to molecular biology. Specifically, in the case of RNA (i.e. RNA-seq), the capacity to rapidly obtain high-quality transcriptome data even from minute amounts of sample broadened its use from model to numerous other organisms with very limited or no molecular data [89]. Moreover, novel variants of RNA-seq, e.g. for exome and single cell sequencing, are still being developed and optimized [267]. While differential gene expression remains the main application of RNA-seq analysis, the same pool of sequence reads can be mapped onto the genome to identify transcribed genomic regions [268]. Hence, they are invaluable for gene prediction and genome annotation and can provide information with respect to the position of coding genes and their exon/intron structure, which is the first and most crucial step in understanding the organization of a newly sequenced genome. Yet, in terms of throughput, experimental set-ups reaching the hundreds and thousands of samples scale are still considered laborious and extremely expensive for most research laboratories. Besides, established RNA-seq methods covering the entire length of transcripts frequently fail to accurately assign reads to gene family members that are the products of recent duplication events and share high sequence similarity [269].

“High throughput 3’ end sequencing” is an emerging RNA-seq method in which only the 3’ end of genes in a sample are converted into cDNA and sequenced (Figure 4.1 and Figure 4.2A-B; [88, 270, 271]. In this approach, early barcoding and multiplexing allows the preparation and sequencing of an extensive number of samples at once, hereby strongly reducing the sequencing cost per sample. Briefly, fragmented poly-adenylated RNA molecules are used as a template for reverse transcription, resulting in a library that contains the 3’ end of all poly-adenylated RNA molecules, including mRNA transcripts as well as micro-RNAs (miRNA) and other noncoding-RNAs [272]; Figure 4.1 and Figure 4.2A-B; Supplementary Method). Consequently, this method is able to reveal the location of 3’ ends of all poly-adenylated transcripts in each sample. Furthermore, to precisely estimate the level of gene expression, oligonucleotides used in this method take advantage of unique molecular identifiers (UMIs, consisting of random nucleotides) to count the absolute numbers of RNA molecules [273]. Furthermore, as only a short region at the 3’ end of genes is sequenced (Figure 4.2C-G), as little as ~1 million reads per sample are required to cover the whole transcriptome of higher eukaryotes [270]; Figure S4.1].

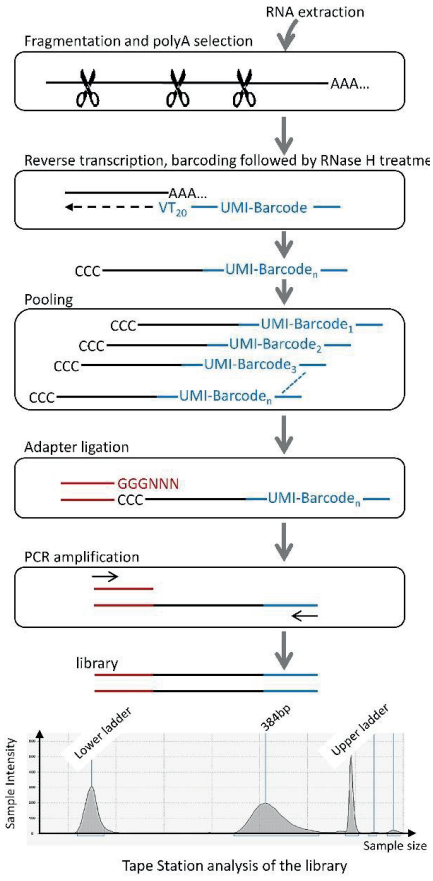


Figure 4.1: ***Workflow of the TranSeq library preparation method.*** *In the TranSeq method, RNA is fragmented into small pieces using divalent cations under elevated temperature and purified using oligo-d(T) magnetic beads. RNA is subsequently used as a template for cDNA synthesis using long barcoded oligonucleotides. Following RNase H treatment, cDNAs are pooled together and ligated to a double stranded adapter followed by PCR amplification to complete the library preparation and examine its quality.*

Few methods for 3' end sequencing of mRNA transcripts have been developed in the past years (e.g. MPSS-DGE and SBS-DGE [274]. These methods could efficiently detect gene expression and revealed that most Arabidopsis and rice genes hold alternative poly-adenylation sites, ~49%–66% which occur upstream of the annotated stop codons. Yet, despite its potential, they were never employed to improve plant genome annotation and better define gene models. Besides, the experimental complexity of these methods, employing restriction enzymes and biotin beads purification, most likely prevented easy and rapid preparation of large amounts of libraries simultaneously. Notably, laborious and lengthy library preparation protocols create more heterogeneity among libraries and often increase the cost of library preparation. One of these methods, termed BrAD-Seq [275], resembles the method presented here. However, in our method, template switching PCR [276] is utilized for the synthesis of the second DNA strand, rather than breathing of cDNA-RNA duplex. The efficiency of BrAD-seq to detect gene expression was at no time compared to that of the "regular" RNA-seq method (i.e. the Illumina RNA-seq procedure) and hence its competence at characterizing gene expression was not assessed. Furthermore, data produced by BrAD-seq was not

employed thus far to better define the genes models at their 3' ends. Additionally, novel methods, particularly for RNA-seq, may often be restricted for use through commercial kits, and this dramatically increases costs and consequently reduces the quantity of users. For these reasons, the use of 3' end RNA-seq methods remained limited, especially in plant research, and have not become a routine in transcriptome and genome analysis.

In this study, we applied both TranSeq, a high-throughput 3' end sequencing method [270], and the established full-length transcript sequencing method (i.e. the Illumina RNA-seq procedure) termed here TruSeq, to characterize the transcriptome of tomato fruit during development. We show that both methods could detect gene expression in a similar manner. We further applied computational analyses to map TranSeq reads to the reference tomato genome and examined if this could significantly improve genome annotation. Our results showed that using TranSeq facilitated the annotation of new transcripts and re-assignment of existing gene models new 3' Untranslated Region (UTR), exon, and intron annotations in the tomato genome. Overall, the datasets generated in this study allowed the improvement of > 45% of the existing tomato gene model predictions and facilitated a new interpretation of the tomato transcriptome. Altogether, TranSeq establishes a new level in throughput for transcriptome sample sets by strongly reducing experiment costs, and thereby shifting experimental set-ups from the current dozens of samples to the hundreds and even thousands in a single experiment. At the same time, TranSeq data provides accurate expression measurements and the potential to detect specific gene family members possessing high sequence similarity and means for significantly improving gene model annotation in the numerous reported 'draft' genomes.

4.4 RESULTS

4.4.1 Sequencing and mapping of TranSeq reads to reference plant genomes

The relatively inaccurate structural annotation of most plant genomes, as well as the limited throughput of samples processed in the standard pipelines for transcriptome analysis, prompted us to assess the use of TranSeq, a high-throughput 3' end sequencing method, in plant transcriptomics assays [270]. TranSeq-based cDNA libraries were prepared from Arabidopsis and tomato fruit tissues. In Arabidopsis, we used seedlings (two weeks old), fully developed siliques, seeds at five development stages, and whole open-flower tissues, while whole fruit at the mature green (MG), breaker (Br), and red ripe (RR) stages were analysed in tomato. In tomato, the same tissues were used to generate TruSeq libraries and to perform full-length transcript sequencing (the standard Illumina method). We then mapped the resulting reads onto the genomes of the corresponding plants. The numbers of sequenced reads generated for each of the libraries are given in Table S4.1. For the TruSeq library, we generated 111,000,000 reads which were filtered for quality to yield 105,000,000 reads (94% of the reads passed the trimming).

Normally, 48 RNA samples were used to prepare one TranSeq library, sequenced in a single lane of an Illumina HiSeq system. This yielded ~35,000,000 reads of which 31,000,000 typically passed the Fastq filter (89.1% reads passed trimming). The reads were then filtered by UMI, i.e. identical reads were not counted unless they harboured a different UMI [273]. Unlike the reads obtained from TruSeq analysis, which are mapped all along the transcripts, the TranSeq reads originate solely from the 3' ends of transcripts (Figure 4.1 and Figure

4.2A-G). Thus, as little as ~2 million reads are sufficient to cover the transcriptome of a higher eukaryote (Table 1 and Figure S4.1).

Table 4.1 ***Comparison between TruSeq and TranSeq throughput and cost performances****:*

Parameters	**TruSeq**	**TranSeq**
Sample to library preparation	2-4 days	1 day
Throughput per experiment	10 - few tens of samples	> 1000 samples
Cost of library preparation per sample	$200-$300	~$8
# of reads for full coverage	~20,000,000	~2,000,000

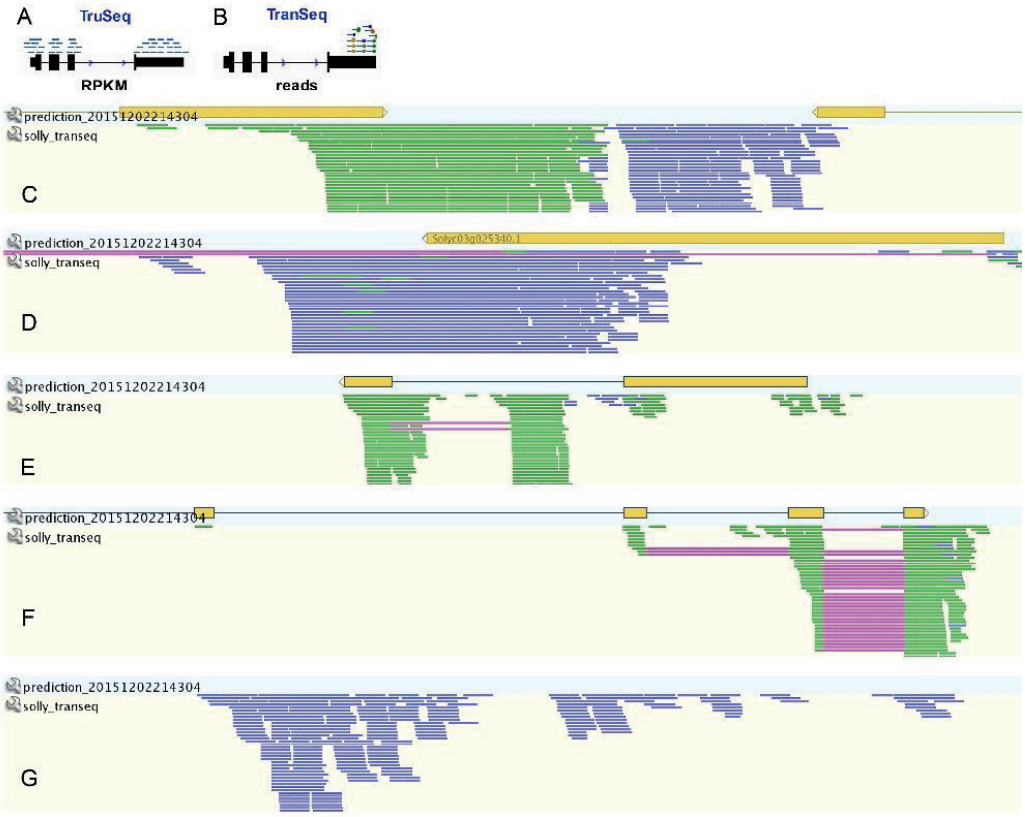


Figure 4.2: ***Mapping of TranSeq reads to the tomato reference genome (ITAG2.4).*** *(A-B) Scheme representing the reads obtained from TruSeq (A) and TranSeq (B) methods, mapped on a typical gene model. (C-D) Examples of expected alignments of reads to the 3' UTR of typical genes. (E) Example of 'orphan reads', which were mapped to a genomic region where there is no gene predicted. (F-G) Example of reads, which were mapped to unexpected locations in gene models (i.e. to exons or introns, rather than to the 3' UTR).*

In an optimally annotated version of a given genome, TranSeq reads would map merely to the 3' UTR of the predicted gene models. We first mapped the corresponding TranSeq reads to the Arabidopsis genome, which is the most advanced annotated plant genome to date. Unexpectedly, only 75% of TranSeq reads were mapped to the 3' UTR of the existing gene models (TAIR10 version), while the remaining 25% were localized

to regions currently annotated as introns, exons, and intergenic regions. Next, we mapped the corresponding TranSeq reads to the reference genome of tomato (ITAG2.4). Likewise, we found that only 54% of the TranSeq reads were mapped to the 3' UTR of the existing gene models, while the rest were localized elsewhere in the genome (Figure 4.2, Table S4.1, and Figure S4.2). Reads mapped to regions where no genes were predicted were referred to as 'orphan reads'. Such low percentages of 3' UTR reads mapping correctly to the 3' end of gene models imply that the genome annotation of these plants can still be considerably improved.

The results described above prompted us to use TranSeq for re-annotating the genome of tomato. To this end, we used a combination of TruSeq and TranSeq datasets (Supplementary Data 1 and Supplementary Data 2), prepared from the same tomato fruit tissues at three developmental stages for de novo gene prediction. Our results suggest an elongation of the 3' UTR of >45% of tomato genes (Figure 4.3, inner track - grey bars). Moreover, the reannotation output not only yielded extended 3' UTR regions, but also refined intron/exon predicted models. In some cases, a 'new exon' was identified (Supplementary Data 1). Lastly, we collected all orphan reads and used them to predict yet unidentified poly-adenylated transcripts (Figure 4.2G). The refined gene models assigned have recently been incorporated into the present version of the tomato genome (www.solgenomics.net).

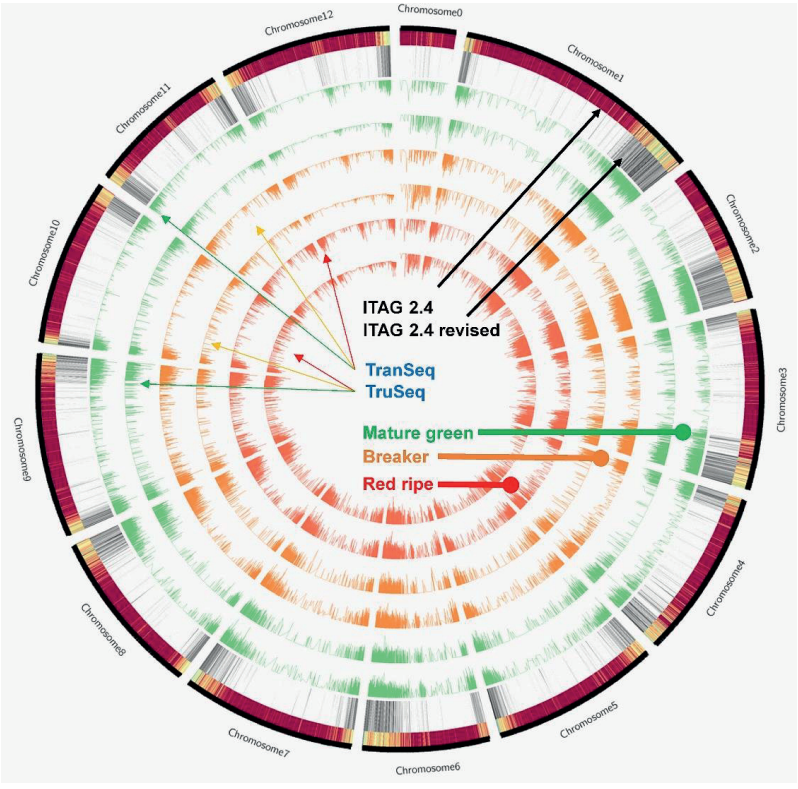


Figure 4.3: ***Original and re-annotated gene models in the tomato genome using TranSeq and TruSeq****. The Circos plot represents the tomato genome (ITAG2.4) divided to 12 chromosomes (outer black lines) and shows gene density (outer track; red and yellow bars represent low and high gene density, respectively) and revised genes (inner track; grey bars), based on TranSeq re-annotated or newly annotated 3' UTR regions. The six most inner tracks outline the expression*

patterns of the shared genes detected by TranSeq and TruSeq methods, at mature green (green track), breaker (orange track) and red ripe (red track) stages. Each chromosome was divided and plotted into 20kb bins.

4.4.2 TranSeq and TruSeq gene expression show similar expression pattern

To determine whether 3' end sequencing is as powerful as the more widely used RNA-seq methods, we compared the TranSeq and TruSeq methods with respect to tomato gene expression (total 9 samples; 3 fruit stages in 3 biological replicates). After filtering for genes with 'above basal level' of expression (sum of counts in all stages is at least 10), 23349 and 17854 genes were detected by the TruSeq and TranSeq, respectively. Among these, 17642 genes were shared by both methods and 72% displayed higher expression correlation with each other than random, as demonstrated by a spearman correlation plot (Figure 4.4) and the pattern of expression in the inner lanes of the circos plot (Figure 4.3, green, orange and red tracks). The differential expression analysis revealed that TruSeq is still more powerful in detection of differentially expressed genes (supplemental table 3), which can be allocated to the difference in library size. Furthermore, Principal Component Analysis (PCA) showed that 36.8% of the variance in expression could be explained by the first component PC1 (Figure S4.4A), which separates between the TranSeq and TruSeq results, and was largely due to the large difference in the averaged level of counts between the two methods. The other components, PC2 (18.9%; Figure S4.4B), PC3 (13.8%; Figure S4.4C) and PC4 (4.9%; not shown) captured the variance between the sampled tissues (i.e. the three developmental stages), and showed a similar trend between the two methods.

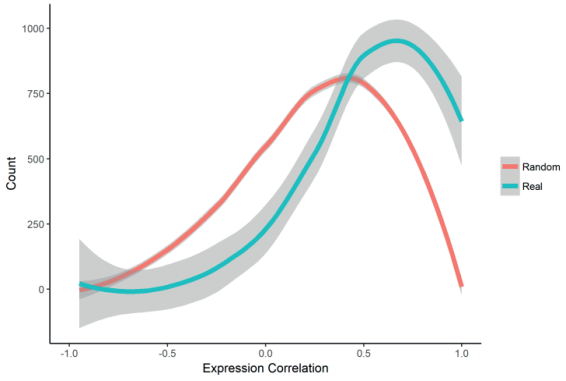


Figure 4.4: ***Co-expression correlation for each gene, between the TranSeq and TruSeq methods.*** *Distribution plot of the correlations (Spearman) of each gene with its counter self in TranSeq vs. TruSeq (blue line). Random distribution of correlations is based on 1000 random samples from the same set (red line).*

4.4.3 TranSeq analysis of gene expression efficiently discriminates between gene family members

Neo-functionalization or sub-functionalization is often the fate of members of a gene family that occurs via accumulation of mutations after duplication [277-279]. Gene family members may therefore share high sequence similarity. Since reads obtained in standard RNA-seq procedures (e.g. TruSeq) can be mapped

equally well to all highly similar regions, these methods often fail to discriminate between recently duplicated (or very similar) gene family members. Due to the typically higher sequence variation in UTR regions, we anticipated that 3' end sequencing methods, such as TranSeq, would facilitate discriminating between gene family members and may even detect differences in their expression. To test this hypothesis, we analysed the TranSeq and TruSeq libraries that originated from the same RNA samples by mapping the reads to the reference tomato genome (ITAG2.4). Gene expression levels were stored in two matrices of nine conditions each. Expression values for either one of the duplicated genes in a specific gene family were extracted [229]. Mean normalized counts were calculated per gene and library separately. We then applied the negative binomial distribution model, using the glimmix procedure in SAS for normalizing each column in the matrices against library size.

We found that the TranSeq method could differentiate between gene duplicates that display low expression level and little sequence divergence (Figure 4.5). Furthermore, we categorized 11,551 genes into 4136 gene families [229], which contained at least two gene members. Using both TruSeq and TranSeq, expression values could be assigned to 3484 gene families, which were further subjected to statistical analysis. Mean normalized counts (i.e. divided by library size) were calculated for each gene for each library separately. In 1418 ortho groups out of 3484, the highest mean over all conditions was found to be a gene from the TranSeq library.

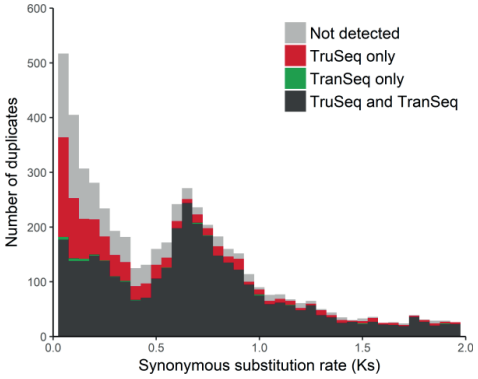


*Figure 4.5: **Sensitivity of gene duplicate detection.** Absolute number of duplicates detected either with TranSeq and/or TruSeq, as function of the number of synonymous substitutions (Ks) between the duplicates. The grey background indicates the total number of duplicates in each Ks bin. The green and red histograms depict genes whose expression could be detected either only in the TranSeq or the TruSeq libraries, respectively. The black histogram depicts number of genes detected both in TranSeq and TruSeq libraries. Duplicates for which Ks >2, were discarded.*

To further corroborate the findings above for both TruSeq and TranSeq, we classified the genes of each gene family to two classes, namely 'highly expressed' (to which the gene with highest mean expression levels in each gene family was assigned), and 'lowly expressed' (to which all other genes in the same gene families were assigned). The analysis is done for each ortho group separately. We then used a t-test to measure the significance of gene expression differences between the 'low' and 'high' classes of gene expression for each developmental stage. We found that in 269 gene families, there was a significant difference between the high and low classes of gene expression ($P<0.05$, across all three fruit developmental stages; Table S4.2). After multiple testing using the Sidak step-down adjusted p-values, the difference between 'high' and 'low'

classes of gene expression were significant in 141 gene families (P<0.05 family-wise significance level in each of the three fruit developmental stages; Supplementary_Plots.pdf).

Furthermore, we found that the duplicates from the 141 gene families, which have a significant differential expression between duplicates, also displayed a significantly lower synonymous substitution rate than expected randomly (Ks; P <0.001; Figure S4.5). The Ks distributions for the 653 gene families (having significant gene expression differences between the 'low' and 'high' classes) was next compared with the complete set of *S. lycopersicum* duplicates (the entire tomato paranome). Notably, we observed that the Ks range of the 141 gene families set largely overlaps with the one associated with the Solanum WGD set (Ks: 0,4-1,0) [21]. Furthermore, block synteny analysis was run for the 653 gene families, using the Chi-square test + multiple testing correction and p-value [229, 242], and revealed the enrichment of these gene families (p-value =1.2e-3; Table S4.2) for duplicated genes originating from the Solanum WGD/hexaploidy event [229]. These findings showed that TranSeq can outcompete TruSeq in the discrimination between gene duplicates and their expression. It also allowed us to demonstrate the significant divergence in gene duplicates originating from a WGD.

4.5 DISCUSSION

In recent years, numerous transcriptome studies performed at the cell-type and single-cell level demonstrated overwhelmingly, in organisms from diverse kingdoms, a vast heterogeneity in the gene expression profiles of seemingly similar cells [280, 281]. Hence, novel methodologies to study this gene expression diversity are invaluable. Nevertheless, they should deal with much larger sample sets as compared to the current norm as well as require a relatively minute amount of RNA per sample [282-286]. The present study demonstrated the power of TranSeq, a high-throughput 3' end sequencing method originally developed for studies in mammalian systems, and its application in transcriptome assays and genome annotation in plants. While current RNA-seq experiments in plants, e.g. using the TruSeq procedure (Illumina, Inc.), will typically comprise a few up to several dozens of samples, the use of TranSeq is expected to increase the sample set per experiment to the hundreds and thousands scale, while keeping the experiment cost reasonably low. Notably, when performing a TruSeq experiment, the transcriptome size of tomato will require a minimum of 20 million reads to allow comprehensive gene expression analysis [287, 288]. The study here demonstrated that as little as 1 to 2 million reads per sample are enough to cover most of the tomato transcriptome when performing a TranSeq experiment. Typically, this allows to analyse up to 48 samples in a single lane of HiSeq (Illumina), which yields about 250M reads, and therefore greatly reduces the cost of sequencing per sample. We expect that the development of more advanced sequencing technologies will allow the analysis of even larger sample sets and further reduce the sequencing costs. In addition, significantly increasing the throughput of transcriptome analysis is expected to require new data processing, quantification and certainly visualization algorithms to allow the extraction of meaningful biological knowledge. Moreover, it is even likely to stimulate breakthroughs in developing advanced technologies to carry out single cell isolation and sampling in plants, which are currently a major limiting factor for executing such experiments [289].

One noteworthy weakness of "standard" RNA-seq methods in which sequence reads cover the entire transcript, is the erroneous assignment of reads in between highly related sequences such as members of

the same gene family [269]. Plant genomes are exceedingly enriched in large gene families that often include very similar genes in tandem gene clusters, that could be part of a duplicated genomic region or spread out across the genome [290]. In such cases (as well as in many polyploid plant species), using the standard full transcript coverage RNA-seq methods to discriminate between expression of genes having substantial sequence similarity, is a major issue of concern. We found that TranSeq, merely generating reads matching the 3' UTRs of transcripts (which are significantly variable even among closely related gene family members), is most effective in determining the expression level of individual, related gene family members. By comparing TruSeq to TranSeq we showed that 75% of the tomato genes displayed a similar pattern of expression in both sequencing methods. Typically, with TranSeq, when examining 151 gene families in tomato, only one gene was expressed to a significant level while the others possessing significant sequence similarity displayed much lower expression.

While the original aim of this study was to increase the throughput of transcriptomics experiments, we also realized the immense potential of the TranSeq approach for whole-genome sequence annotation. To date, more than a 100 de novo-sequenced plant genomes have been publicly released. However, many of them suffer from extensive fragmentation and poorly defined gene models [291]. Following this work, we propose a complementary approach that takes advantage of two RNA-seq methods for the (re)evaluation of gene models. The currently standard RNA-seq procedure to obtain complete transcript sequences was combined with a 3' end sequencing method to re-annotate the tomato genome. It appeared that a large portion of the tomato genome gene models was miss-annotated at its 3' end. Our approach allowed the correction of more than 45% of the tomato gene models (version ITAG2.4), by assigning them either a longer 3' UTR, or an additional exon at their 3' end, or various extensions of their 3' UTR. Notably, TranSeq could also discriminate between transcripts that display multiple polyadenylation sites while TruSeq was not able to differentiate between those. Notably, proper annotation of 3' UTRs is of great importance to understand gene control, e.g. the selection of polyadenylation sites and 3' UTR length may result in different intrinsic stabilities of a given transcript [292].

Despite the multiple strengths of using the TranSeq method portrayed above, it is important to note the restricted use of this method to organisms having a minimal quality of whole-genome sequence annotation. Consequently, its use in settings such as gene expression in a natural diversity sets including transcript profiling of entire Genome Wide Association Studies (GWAS) population will likely be reserved to those species with de novo sequenced genomes. Apart from the projected impact of the TranSeq method on the resolution of transcriptome studies, we anticipate that this method will be integrated into gene space annotation strategies of newly sequenced plant genomes.

4.6 MATERIALS AND METHODS

4.6.1 Plant material and sequencing libraries preparation

Two weeks old Arabidopsis Col-0 seedlings, whole open flowers, seeds at five development stages (12, 14, 16, 18, and 21 days after pollination), fully developed siliques, and entire tomato fruit at the mature green, breaker and red ripe stages (cv. MicroTom) were frozen in liquid nitrogen prior to RNA extraction and grinded for RNA extraction using the TRI-reagents method [293]. A detailed protocol for the preparation of TranSeq

libraries (single end 60 bp reads) is provided in the Supplementary Method and the overall workflow of the method is presented in Figure 4.1. The protocol for the preparation of TruSeq libraries (single end 60 bp reads), was described in [294].

4.6.2 Mapping of sequenced TruSeq and TranSeq reads to the reference genomes

For TruSeq and TranSeq data was mapped using Bowtie2 v2.2.5 [295] and TopHat v2.0.13 [296] to the tomato (ITAG 2.4) and Arabidopsis (TAIR 10) reference genomes. Once mapped, reads were visualized with the GenomeView tool [297]. For TruSeq HTseqcount was used and for TranSeq an in-house UNIX shell script that invokes other PERL and Java scripts, was used for counting and quantifying gene expression (see Supplementary Method).

4.6.3 Genome (re)annotation

The new annotation was generated using a combination of extrinsic evidence and the Eugene software able to include extrinsic information while building gene models. This approach is a slightly modified version of the pipeline used for the original ITAG annotation. As a final step we also included EvidenceModeler (EVM; [298], which checks Eugene predictions for support and is able to apply changes to gene models to augment the general support for gene-models using the provided alignments. As extrinsic protein data, we used data from ITAG annotations from both *S. lycopersicum* as *S. tuberosum* (www.solgenomics.net), after being cleaned from genes with a transposable element related functional description or short (<300nt) hypothetical genes. Additionally, we run the pipeline a few times, each time including the 'best' predicted tomato proteins from the previous run in a way that will provide enough evidence to properly predict a member of a gene-family, and this predicted gene will serve as the best homologue available to assist with the prediction of the other members of the same gene-family. The 'best' predicted genes are those that are (+/- 10%) of comparable length with the best-blast-hit against a reference protein set (therefore we compared the gene models to Arabidopsis TAIR10, since it is the best annotated plant species). Besides the protein support, we also included transcript data including ESTs collected from NCBI, RNA-seq (TruSeq) junctions (reads spanning introns) and transcript contigs (from assembled TruSeq RNA-seq) cut back in contigs of maximum 500nt. Cutting back the RNA-seq contigs to 500nt allowed reducing the chimeric contigs with misleading mappings. The TranSeq RNA-seq was independently assembled into contigs, and the contigs used as full length. The aim of this last set was to extend the gene-model's UTRs in a most reliable way. All the above data was mapped on the SL2.5 assembly, that was masked using Rebase20.1 added with a custom repeat library (RepeatModeler, default parameters), and used to produce gene-models using Eugene. The extrinsic evidence was then converted into GFF format appropriate for including in EVM for further possibility of local improvements.

4.6.4 Gene expression profiles of gene families in tomato.

We analysed libraries of TranSeq and TruSeq reads that originated from the same RNA samples (3 stages in 3 replicates) and stored in two matrices (of nine conditions each). The expression matrix was transformed using rlog transformation [299]. The PCA was performed using the prcomp function in R. Spearman expression correlation of the genes expressed in both TranSeq vs. TruSeq were compared to random

distribution was based on 1000 random samples from the same set. Since total library size was much higher in TruSeq than in TranSeq, we normalized the read counts by library size (to inflate the read counts in TranSeq). The negative binomial distribution was used to model the count data with the library size as offset variable using the glimmix procedure in SAS. Fixed effects were expression and condition and their interaction. Simple test of effects was calculated between the low and high expression level for each condition. Multiple testing to obtain adjusted p-values (p-value < 0.05), for testing for significant difference between 'high' and 'low' expression family-wise significance level in each of the three conditions was done with the Sidak stepdown function in SAS. Gene families/gene and the corresponding *S. lycopersicum* duplicates with Ks values were taken from [229]. Correlation analysis and plotting of the frequency distributions of the correlation coefficient (r) values of each pairs of duplicated genes (compared to the r-value obtained from random pairs of genes), was performed in R (version 3.1.1).

4.6.5 Computational resources

All data processing and data analysis was done using a combination of Python and R scripts. The read mapping and counting pipeline was run for all samples in parallel on a computing cluster with Linux computing nodes (2.4GHz, Intel). All samples were run individually on 4 nodes with each 4GB of memory. The running time for TranSeq samples was between one and two hours. For TruSeq the running time was between 2 and 5 hours.

4.7 SUPPLEMENTAL DATA

All supplemental files can be found on:

https://floppy.psb.ugent.be/public.php?service=files&t=2cc513eced261bef2ebdd8b4c3b042d6

4.7.1 Supplemental figures

Figure S4.1: ***Filtering TranSeq library reads in tomato****. Number of sequence reads that passed each filter during TranSeq data analysis for all 9 tomato fruit samples analysed in this study at the mature green (MG), breaker (Br), and red ripe (RR) stage.*

Figure S4.2: ***TranSeq and TruSeq reads map to different genomic features.***

Figure S4.3: ***Examples of TranSeq (green) and TruSeq (blue) based re-annotation, extends and modifies tomato gene models in ITAG2.4 (the "prediction" track).*** *(A) Mapping of TranSeq and TruSeq reads on the extended UTR gene model. (B) The SL2.4_UTR is the extended gene model of Solyc05g012040 (showing a new intron gained).*

Figure S4.4: ***Principal Component Analysis (PCA) of reads obtained from tomato TranSeq and TruSeq data.*** *Mature green stage (green track), breaker stage (red track) and red ripe stage (blue track). The first component 36.8%(PC1; panel A) separates TranSeq and TruSeq due to the inherit variance in averaged level of counts between the two methods. The other components PC2 (18.9%; panel B), PC3 (13.8%; panel C) capture the variance between the sampled tissues.*

Figure S4.5: ***Density plot of the synonymous substitution rate.*** *This density plot shows the Ks distribution for the duplicate genes from the gene families which have a significant difference between high and low expression (red) and all the other duplicates (grey) in S. lycopersicum. The median Ks for both sets is indicated with a dashed line. Duplicates for which Ks >2, were discarded. The purple background indicates the Solanaceae WGD peak area (based on [21]).*

4.7.2 Supplemental tables

Table S4.1: ***TranSeq and TruSeq read mapping to the tomato reference genome. Reads where mapped twice (once to the original tomato genome named- ITAG2.4_gene_models; and once to our extended gene models named – SL2.5extUTR_intergenic – column B)****, and categorized into five different genomic features: CDS, 5' UTR, 3' UTR,*

intergenic regions, and introns (column C). In columns D-G we summarize the number of expressed genes with at >0, >5, >10 and >20 counts for each reference genome and sequencing method.

Table S4.2: ***List of genes that displayed high expression levels across all samples, comparing to other members of their gene family measured from TranSeq libraries****. Expression levels, of the rest of the genes in the families, measured from TruSeq libraries were constantly low.*

Table S4.3: ***Comparison of expression and differentially expressed genes in TranSeq and TruSeq.*** *The first part contains the number of detected genes in the different conditions with the overlap between them, followed by the number of differentially expressed genes (P<0.001). The tables on the bottom show the number of duplicates which are either, both, only one or are not detected using both techniques.*

4.7.3 Supplemental data

*Supplementary Data 1****. TruSeq gene expression tomato dataset*** *(FPKM).*

Supplementary Data 2. ***TranSeq gene expression dataset*** *(counts).*

5

FUNCTION, DYNAMICS AND EVOLUTION OF NETWORK MOTIF MODULES IN INTEGRATED GENE REGULATORY NETWORKS OF WORM AND PLANT

"...Data itself... was tolerable. It was the constant nerve-web-expanding pain of context that would kill him."

Dan Simmons

5 NETWORK MOTIF MODULES IN INTEGRATED GENE REGULATORY NETWORKS OF WORM AND PLANT

This chapter is a manuscript submitted to Nucleic acid research

Jonas Defoort [1,2,3], Yves Van de Peer [1,2,3,4] & Vanessa Vermeirssen [1,2,3]

[1] Department of Plant Biotechnology and Bioinformatics, Ghent University, 9052 Ghent, Belgium

[2] VIB Center for Plant Systems Biology, VIB, 9052 Ghent, Belgium

[3] Bioinformatics Institute Ghent, Ghent University, 9052 Ghent, Belgium

[4] Department of Genetics, Genomics Research Institute, University of Pretoria, Pretoria 0028, South Africa

5.1 ABSTRACT

Gene regulatory networks (GRNs) consist out of different molecular interactions that closely work together to establish proper gene expression in time and space. Especially in higher eukaryotes, many questions remain on how these interactions collectively coordinate gene regulation. We study high quality integrated GRNs consisting of undirected protein-protein, genetic and homologous interactions, and directed protein-DNA, regulatory and miRNA-mRNA interactions in the worm *Caenorhabditis elegans* and the plant *Arabidopsis thaliana*. Our data-integration framework integrates interactions in composite network motifs, clusters these in biologically relevant, higher-order topological network motif modules, overlays these with gene expression profiles and discovers novel connections between modules and regulators. Similar modules exist in the integrated GRNs of worm and plant. We show how experimental or computational methodologies underlying a certain data type impact network topology. Through phylogenetic decomposition, we found that proteins of worm and plant tend to functionally interact with proteins of a similar age, while at the regulatory level TFs favour same age, but also older target genes. Despite some influence of the duplication mode difference, we also observe at the motif and module level for both species a preference for age homogeneity for undirected and age heterogeneity for directed interactions. This leads to a model where novel genes are added together to the GRNs in a specific biological functional context, regulated by one or more TFs that also target older genes in the GRNs. Overall, we detected topological, functional and evolutionary properties of GRNs that are potentially universal in all species.

5.2 CONTRIBUTION

- Performing the research together with Vanessa Vermeirssen
- Designing and performing analyses together with Vanessa Vermeirssen
- All Figures and tables
- Writing the manuscript together with Vanessa Vermeirssen

5.3 INTRODUCTION

In eukaryotic organisms, differential gene expression is a tightly controlled process that governs developmental, physiological and disease processes. At the level of transcription, specific transcription factors (TFs) bind DNA in order to activate or repress the expression of a gene. MiRNAs repress gene expression post-transcriptionally by interacting with complementary sequences located in the 3'UTR of their target mRNAs. Many molecular interactions, in which TFs and miRNAs are key players, closely work together in order to establish proper gene expression in space and time [300, 301]. In addition to binding DNA at specific regulatory sites in the genome, several TFs influence transcription through protein-protein interactions, either because they bind DNA as homo- or heterodimers, or because they require interaction with cofactors, chromatin modifying factors or the basal transcription machinery [302]. In addition to these direct physical interactions, other molecular interactions have an indirect impact on gene regulation. Genetic interactions, in which two mutations have a combined phenotypic effect not exhibited by either mutation alone, reveal functional linkages in gene regulatory circuits [303]. Together with paralogous interactions, which occur frequently between TFs and miRNAs, since duplication events significantly contributed to their evolutionary expansion [304, 305], they can also pinpoint redundancy in gene regulation. Regulatory interactions between TFs and target genes are identified by expression profiling in organisms with perturbed TFs and describe both direct and indirect influences of these TFs on gene expression.

While we understand the biological consequences of single data types, we are just beginning to explore how different interaction types together influence gene regulation. E.g. co-expressed genes and genes encoding interacting proteins tend to be regulated by common TFs [306, 307]. Synthetic genetic interactions are more likely to occur between homologous genes, although large gene families complicate the identification of digenic interactions [308]. Genes encoding TFs that control miRNA expression have a higher chance to be post-transcriptionally repressed by the miRNA [309]. Furthermore, genes coregulated by miRNAs are less functionally linked than genes coregulated by TFs [310]. Therefore, different types of molecular interactions provide complementary insights into gene regulation and cell function, expressing the need for data integration [311].

Gene regulation can be studied in gene regulatory networks (GRNs), which map the interactions between TFs and their target genes at a systems level [312]. Taking into account different types of molecular interactions that specify regulatory inputs, generates integrated GRNs. Network motifs, which are defined as patterns of interconnections occurring significantly more often than in randomized networks, have been regarded as the basic building blocks of complex networks [313]. More specifically, the feed forward loop (FFL), which with positive regulations acts as a signal persistence detector, is the most prominent motif in GRNs of Escherichia coli and Saccharomyces cerevisiae [314-316], and also in higher eukaryotes like human [317, 318]. Similarly, integrated GRNs can be characterized by composite network motifs, which are subgraphs of which the edges can represent different interaction types, e.g. TF complexes regulating a common target gene and the transcriptional coregulation of interacting proteins [78]; miRNA-TF feedback loop and miRNA-mediated FFL [309, 319-321]; TF-regulated kinate motifs and interacting kinates motifs [322], and CoRePPI motifs considering coregulation of protein-protein interactions by TFs and miRNAs [323]. Hence, studying composite network motifs in integrated GRNs has already revealed novel topological structures with biological implications that cannot be deduced from single interaction type networks.

The relation between motif type and biological function has been debated [324-326]. Detailed information about a motif's signal integration logic, i.e. binding site affinities and molecular interactions of the regulatory TFs, is necessary for a complete understanding of the motif's function. In addition, not only network motifs, but the higher topological patterns into which they cluster, determine biological function. In GRNs of E. coli and S. cerevisiae, networks motifs such as the FFL aggregate into homogeneous motif clusters, mostly multi-output FFL generalizations, that largely overlap with known biological functions [327, 328]. Also, in integrated GRNs of S. cerevisiae, composite network motifs cluster together in recurring interconnecting patterns that could be tied to specific biological phenomena such as for instance in the regulonic complex theme wherein a TF regulates multiple members of a protein complex, both TF and protein complex tend to be involved in the same biological process and complexes of related function are often connected to the same TF [326, 329, 330]. A single composite network motif can aggregate into topologically distinct motif clusters e.g. a motif composed of a transcription regulatory interaction where regulator and target both physically interact with the same protein, can cluster either into a "regulonic star", where multiple targets of a TF interact with the same feedback mediator, or a "regulatory interacting double-star", consisting of a regulator-target pair that share a common set of partners in the protein interaction network, which usually belong to a regulatory protein complex [330]. Diverse complex networks exhibit rich higher-order organizational structures that are exposed by clustering based on higher-order connectivity patterns and hence provide biological contextualization [331].

The current GRNs are the result of evolution during millions of years. Interaction rewiring and integration of novel genes is an important step in this process. Novel genes originate through partial or full duplication of existing genes followed by divergence, incorporation of mobile elements, gene fission and fusion, and de novo gene creation from non-coding sequence [332]. Through phylogenetic analysis, the age of genes can be assigned based on the oldest common species with an ortholog [333]. Studies focusing only on protein-protein interaction and coexpression networks in different eukaryotic species revealed that the majority of young genes are incorporated in the periphery and slowly acquire more interactions and functions [334-337]. Novel genes gain interactions and functions faster than duplicated genes [338]. In addition, genes tend to interact more with proteins of the same age in protein-protein interaction networks of yeast [338, 339] and human [334] and in coexpression networks of *A. thaliana* [337]. In yeast it has been shown that proteins with the same age tend to clusters into motifs, while proteins from different age groups tend to avoid motif formation [339]. Based on the observations in yeast protein-protein interaction networks, modelling approaches have tried to mimic network evolution [340, 341]. The best results were obtained with the network motif model where network motifs or protein clusters instead of single proteins are incorporated into pre-existing networks over evolutionary time [341]. Overall, studies on GRN evolution have mainly been limited to protein-protein interaction networks in unicellular organisms.

In eukaryote organisms, the main sources of duplicates are SSD and WGD. In *C. elegans* there is a high rate of SSD, mostly single gene duplications. These SSDs are frequently partial or do not have all regulatory sequences from their original sequences [342]. In plants, next to SSD, there are also WGDs. These can either be the result of interspecific or intraspecific hybridizations which lead to multiple genomic copies (polyploidy). WGDs are very abundant and an important source of duplicates in a wide range of plant species [16]. In *A. thaliana*, there are four or five ancient WGD events described [19, 20], two of which are located

between the Brassicales and the Brassicaceae age groups, one between the flowering plants and the split between eudicots and monocots, and one or two between the origin of seed and flowering plants. After duplication, most of the duplicates get lost [229]. However, many WGD-duplicates evolve slower than SSD-duplicates in terms of divergences of sequence [227], expression [226], protein interaction partners [63] and regulatory connections [343].

Here, we developed a data-integration framework starting from different types of interaction networks, over composite network motifs, to network motif modules, which could be dynamically investigated through integration of expression profiles and topologically interpreted in a 'superview' analysis (Overview pipeline in Figure 5.1). We studied two model organisms that are different from a structural, physiological and evolutionary point of view, i.e. the multicellular worm *C. elegans* and the flowering plant *A. thaliana*, for which many data are available. We learned that different molecular interactions interrelate in biologically relevant network motif modules to generate a coordinated response in gene regulation. Our approach enabled us to show the advantages and pitfalls of data integration of multiple data types and different experimental methodologies on the motif and motif network module level. Next, using phylogenetic decomposition, we investigated how novel genes are incorporated in these networks. We also found that worm and plant proteins prefer to interact with proteins of a similar age. For protein-DNA interactions on the other hand, we found in both species that regulatory TFs favoured to bind to older or of similar age target genes. In network motifs, undirected interactions preferentially took place between age homogeneous proteins, while directed interactions were inclined to be age heterogeneous. These preferred age patterns in the motifs were favourably incorporated in the network motif modules. Modules were mostly composed out of genes from the evolutionary age groups containing the most genes i.e. Eukaryota, Eumetazoa and Caenorhabditis in worm, and Green and Land plants in plant. Hence, especially in *A. thaliana*, younger genes were more inclined to attach to modules mostly composed out of older genes instead of forming modules on their own. Modules with directed interactions were only age homogeneous in the oldest evolutionary age groups or there were none. This leads to a model where novel genes are added together to the GRNs in a specific biological functional context, regulated by one or more TFs that also target older genes in the GRNs.

5.4 RESULTS

5.4.1 High quality integrated GRNs in worm and plant feature hubs and modularity

Given that gene regulation is influenced by different physical and functional molecular interactions, we integrated high quality directed protein-DNA (D), regulatory (R) and miRNA-mRNA (M), and undirected protein-protein (P), genetic (G) and homologous (H) interactions to obtain a holistic view on gene regulation (Figure 5.1). The data contained only experimentally validated interactions, except for miRNA-mRNA interactions, where computational predictions complemented experimental interactions. Homologous relationships were also inferred through computational analysis (Methods). The integrated GRNs of *C. elegans* (Cele) and *A. thaliana* (Atha) contained respectively 43,943 and 89,679 molecular interactions, distributed over the different molecular interaction types as depicted in Figure 5.2. There is limited overlap between the different types of interactions in both GRNs (Suppl. Fig. 1). In the *A. thaliana* GRN, protein-DNA interactions and transcription regulatory interactions are merged in the same D data type due to indistinctness in experimental origin or overlap between the two types of interactions: at least 4236 interactions are both physical protein-DNA and transcription regulatory interactions. Like most biological networks, these networks are scale-free and feature hubs, highly connected proteins in the undirected networks and regulators with many targets in the directed networks (Supplementary Data, Suppl. Table 1, Suppl. Fig. 2) [344, 345]. Many medium-degree nodes have a higher clustering coefficient than expected from the power-law fit (Supplementary Data, Suppl. Fig. 3). Hence, they differ from hierarchical scale-free networks and exhibit an extra modularity than the one centred on hubs [330]. The overall clustering coefficients of the protein-DNA and protein-protein interaction networks of *C. elegans* are 2 to 10 times higher than those of *A. thaliana.* Hence, the worm integrated GRNs are smaller (edge to node ratio Cele 3.4 versus Atha 4.3) and more likely to form clusters.

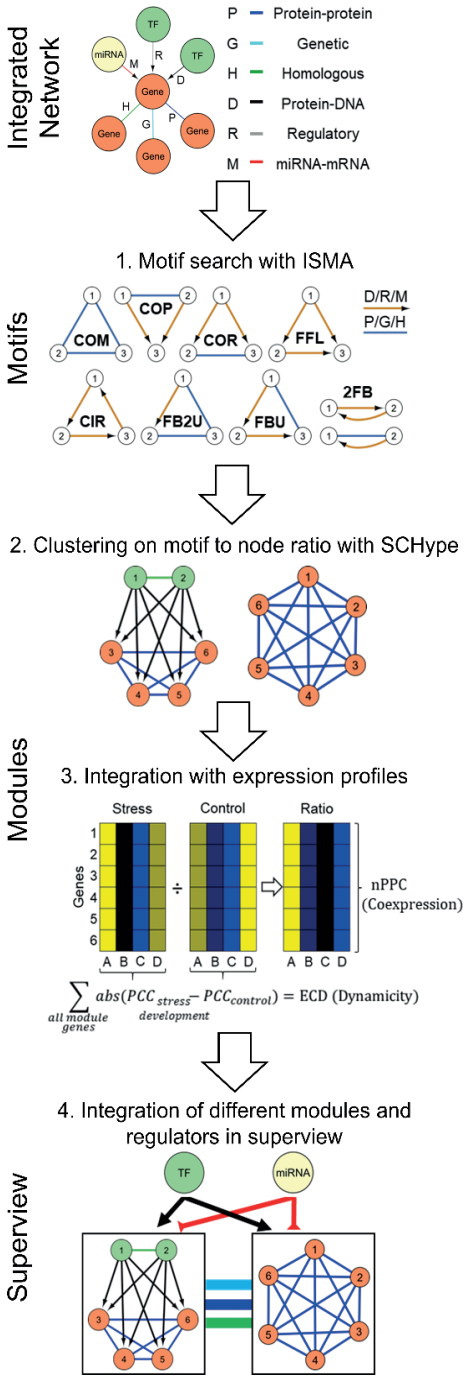


Figure 5.1: ***The data integration framework to study integrated GRNs****. In the first step, molecular interaction data were gathered from multiple sources: protein-protein (P), genetic (G), homologous (H), protein-DNA (D), regulatory (R) and miRNA-mRNA interactions. In the motif step, all possible 2-node and 3-node motifs were searched with ISMA, the Index-based Subgraph Matching Algorithm that conducts a fast and efficient motif search through carefully selecting the order in which the nodes of a query motif are investigated. We grouped the motifs in 8 categories (complex motif (COM), feed forward loop (FFL), co-pointing motif (COP), co-regulated motif (COR), circular feedback motif (CIR), feedback undirected motif (FBU), feedback 2 undirected motif (FB2U) and two-node feedback motif (2FB)) and named them ABC according to the interactions A between node 1 and 2, B between node 1 and 3, and C between node 2 and 3. For directed edges, if the direction is reversed e.g. interaction A between node 2 and node 1, a small case letter is used e.g. motif aBC. In the module step, motifs were clustered with SCHype, which is a spectral hyper-edge clustering algorithm maximizing the hyper-edge (i.e. motif) to node ratio. In the dynamic module step, for each module, coexpression was evaluated by the average Pearson Correlation Coefficient (nPCC) and for A. thaliana dynamicity was assessed by the Expression Correlation Differential Score (ECD). In the superview step, modules were integrated with other modules and regulating transcription factors and microRNAs. This integration was based on statistical enrichment by comparing the observed versus expected interactions through comparison with random modules of the same sizes (Methods).*

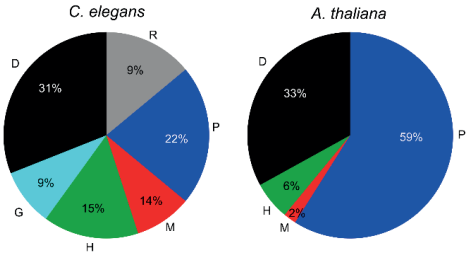


Figure 5.2: ***Proportions of the different types of molecular interactions within the integrated GRNs of C. elegans and A. thaliana****: P – protein-protein, H – homologous, G – genetic, D – protein-DNA (and/or transcription regulatory in the case of A. thaliana), M – miRNA-mRNA and R – transcription regulatory interactions. The worm integrated GRNs contained 43943 interactions between 845 TFs (92% of all), 172 miRNAs (67% of all) and 12095 protein-coding genes (67% of all). The plant integrated GRNs encompassed 89679 interactions between 1519 TFs (88% of all), 174 miRNAs (41% of all) and 19001 protein-coding genes (69% of all).*

5.4.2 Different composite network motifs form the basic building blocks of integrated GRNs

As a first step of our data integration framework (Figure 5.1), we searched for possible 2-node motifs using a customized Perl script and 3-node composite network motifs using ISMA (Index-based Subgraph Matching Algorithm) [346]. These 2- and 3-node motifs are the elementary building blocks of many higher-order motifs. We detected respectively 40 and 14 different 2- and 3-node network motifs that occurred 50 times or more in the GRNs of worm and plant (Figure 5.4, Suppl. Table 2, Suppl. Fig. 4). The composite network motifs were grouped in 8 motif types, all of which were present in both species (Figure 5.4): complex motifs (COM), which represent combinations of all undirected interactions; co-pointing motifs (COP), where two interacting regulators (e.g. dimers or homologs) regulate the same gene; co-regulated motifs (COR), where one regulator controls two interacting genes; feed forward loops (FFL), where a regulator regulates a target gene directly and indirectly through another regulator; circular feedback motifs (CIR), where regulators act upon each other through a feedback loop; feedback undirected motifs (FBU), where two directed interactions in a cascade are connected by one undirected interaction; feedback 2 undirected motifs (FB2U), which combine two undirected interactions and one directed interaction; and two-node feedback motifs (2FB), which couple a directed edge with an undirected edge [77]. The name of the motifs, e.g. RPD, determines the motif: the first letter refers to the left edge from the top node, the second letter refers to the right edge of the top node and the third letter refers to the basal edge in the motif from left to right. A lowercase letter indicates reversal of the directed edge direction.

The higher presence of some motifs in one species as compared to the other can be attributed to the characteristics of the underlying data and methodologies (Figure 5.3, Suppl. Fig. 4). The respective fivefold and twofold higher abundance of P and D data in plant compared to worm generally resulted in higher numbers of P and D containing motifs in plant e.g. PPP (10x), DDD (2x), PDD (3x), DDP (4x), DDM (4x). HHH-motifs are only found in worm, since homologs in *C. elegans* are composed of direct BLAST results, while homologs in *A. thaliana* are based upon gene trees of gene families (Methods). In *C. elegans*, extensive yeast one-hybrid (Y1H) and yeast two-hybrid (Y2H) mapping between TFs led to more widespread TF-TF interactions [347] and hence higher motif counts for the motifs DdD (30x), DmD, DD (19x) and DP (6x). The threefold higher abundance of M data in worm, as well as the large fraction of experimental data in there,

compared to plant, mostly produced higher numbers of M containing motifs in worm e.g. MMD (7x), HMM (32x), MMP (3x) and DM. The higher numbers of MMH (8x), as well as HDD (5x) and DDH (3x) in plant are also possibly caused by the WGD events in *A. thaliana*, where upon duplication of a target gene or TF, also the regulatory edge is duplicated. In both species, specific motifs largely overlap due to overlaying P and D, intersecting P, G and H, and bidirectional D interactions (Supplementary Data, Suppl. Fig. 5): within COP and FB2U, between FBU/FFL, COP/FFL, F2BU/FBU, FBU/COR, FB2U/COR and FFL/CIR. The overlap between FFL and CIR motifs (DDD/DdD) indicate that in FFLs containing only TFs, the final targeted TF transcriptionally feedbacks on the top regulatory TF. A particular difference between plant and worm integrated GRNs is that homologous plant TFs targeting the same genes tend to physically interact more both through protein-protein and protein-DNA interactions.

Typically, the presence of network motifs is evaluated by network motif enrichment. Network motif enrichment was calculated compared to 1000 randomized networks with preserved degree distributions, as is usual done (Methods) [320]. All motif types had at least one network motif enriched in the GRNs (Suppl. Table 2, Suppl. Fig. 4). We found network motif enrichment to be biased towards network topology, which is inherently connected to the experimental methodology (Supplementary Data). As a predominant example, since we integrated chromatin immunoprecipitation (ChIP) and Y1H for D type physical protein-DNA interactions, we observed an enrichment of FFL (DDD) only in the ChIP data and not in the Y1H data for both species (Figure 5.4, Suppl. Table 3). Overall, the net result is a lack of enrichment of the FFL in the integrated GRNs of worm and plant (Figure 5.4, Suppl. Fig. 4). Accordingly, enrichment of the FFL was reported in several studies with a similar randomization methodology, most of them using ChIP data or genome-wide target gene prediction based on conserved TF binding sites for the directed edges [78, 313, 320, 348]. The latter are TF-centred GRN approaches, which result in genome-wide networks at the target gene level with low interconnectivity and few TF hubs. Y1H, on the contrary, is a gene-centred approach, leading to smaller networks with a higher interconnectivity distributed over more TFs and many target gene hubs [349]. Therefore, these data are complementary in the construction of GRNs. This differential network motif enrichment can be mainly attributed to the randomization strategy that preserves the degree distribution, but at the same time limits the randomization in a network topology created by Y1H (Suppl. Table 4). In addition, we also observed that network motifs can be created by the integration of different experimental methodologies for a certain data type. As an example, extra circular feedback motifs DdD originated from the integration of ChIP and Y1H protein-DNA interaction data (Suppl. Table 3). Also, preferential interaction patterns between TF hubs in ChIP and target gene hubs in Y1H emerged in the randomized networks upon data integration, further disturbing the network motif enrichment (Suppl. Table 5).

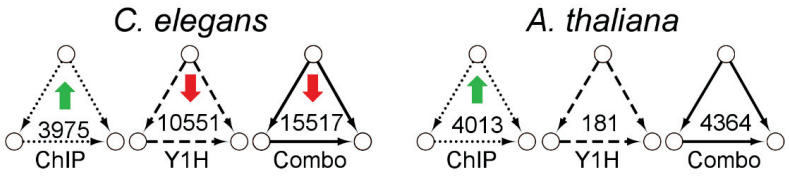


Figure 5.3: ***Differential FFL enrichment between ChIP and Y1H data: the FFL motif is only enriched in the ChIP data of both C. elegans and A. thaliana****. Number of FFL motifs (DDD) in ChIP, Y1H and the combined D data of C. elegans and A. thaliana. Significant over- or underrepresentation as compared to 1000 randomized networks (p-value = 0.05) is indicated by green or red arrow respectively.*

Hence, the experimental or computational methodology that generates a certain data type can exert an impact on the network topology, and more specifically on network motif enrichment. The presence of specific network motifs and their aggregation might therefore be a better indicator of biological functionality of a network than network motif enrichment. Moreover, network motif aggregation is independent of the overrepresentation of network motifs in the network [330]. On top of that, the integration of different data methodologies creates network motifs that would be absent in a single data source network, further indicating that different methodologies are complementary for a given data type to obtain a systems view on gene regulation.

5.4.3 Network motifs aggregate into functional network motif modules

Through a general data-integration framework based on spectral clustering of hypergraphs [350], we investigated the aggregation of motifs into higher order topological structures in the integrated GRNs that represent biological and/or regulatory entities (Figure 5.2). The clustered structures, from now on referred to as network motif modules, can be composed of one type of motif or a combination of different motifs. We classified the modules in 7 different cluster types, depending on which 3-node motifs were clustered together (Methods) (Figure 5.4). In addition, we also clustered all 2- and 3-node motifs together. We functionally annotated the modules with GO Biological Process and investigated dynamic modules by integrating expression profiles from respectively a developmental and abiotic stress expression profile compendium for *C. elegans* and *A. thaliana* [351-353] (Suppl. Tables 7, 8 and 9). Here, we calculated two measures: the average Pearson Correlation Coefficient (nPCC) as a measure of coexpression and, for *A. thaliana*, the Expression Correlation Differential Score (ECD), which highlights modules specific for a stress condition as compared to control conditions (Methods). The number of different network motif modules obtained can be found in Suppl. Table 6.

The first cluster type is the complex module (COMc), generated by clustering the COM motif type (Figure 5.4). Complexes are built out of functionally associated genes, linked through physical protein-protein interactions or/and functional genetic or homologous interactions (cfr. protein complex theme [329]). Proteins in these modules usually show coherent coexpression patterns across conditions. In addition to clusters consisting of only one interaction type, we found clusters composed of members of a protein complex genetically interacting with the same set of proteins (Cele COMc 104 e.g.), since members of a given protein complex or biological process often have common synthetic genetic interaction partners [354]. Furthermore, we observed network motif modules consisting of homologs physically interacting with the same proteins. This is in agreement with the fact that gene duplicates initially have the same interaction partners, before divergence or loss. For instance, in Cele COMc 70, functioning in ubiquitin-dependent protein catabolism, we observed a star-like configuration of protein interaction partners around the homologs cdc-48.1 and cdc-48.2. Both paralogs are also linked by genetic and protein-protein interactions and show a similar expression pattern. Their human homologs suppress the aggregate formation of a Huntington polyQ repeat [355] (See also Supplementary Data for further examples).

The second cluster type is the co-regulated module (CORc), consisting of clusters of the COR motif type, and represents co-regulated functionally associated proteins (cfr. regulonic complex theme [329]). This cluster type adds a transcriptional (e.g. Cele CORc 8) or posttranscriptional (e.g. Atha CORc 14) regulatory layer to a

complex module (Figure 5.4). Here, we also observed star-like configured co-regulated heterodimers (Supplementary Data).

The third type, the co-pointing module (COPc), represents interacting regulators that share a group of targets (cfr. co-pointing theme [329]). We found homologs, heterodimers and protein complexes regulating a set of genes. In the HMM module Cele COPc 61, we observed strongly co-expressed genes involved in axon extension. We also observed interacting protein complexes that combine two groups of functionally associated proteins with regulatory interactions between them. In Atha ALLc 70, a module that combines COP, COR, and COM motifs and changes dynamically upon oxidative stress, the heterodimer PIF3-HY5 targets a number of physically interacting anthocyanin biosynthetic enzymes (Figure 5.4). Finally, we detected homologous signalling pathways like in Cele COP 193 (Supplementary Data).

The fourth type, the feed forward loop module (FFLc), consists only of regulatory links and enables universal information processing and hierarchical regulation [330]. We found the feed-forward theme, where one TF regulates another one and both of them regulate a common set of target genes [329] (e.g. Atha FFLc 30), and extensions to this theme with more regulatory layers or combinations of transcriptional and posttranscriptional regulation (e.g. Atha FFLc 48) (Figure 5.4).

The fifth type, the circular feedback module (CIRc), has not been described before [329, 330]. Here, transcription and/or posttranscriptional regulatory links feed back into one another generating intrinsically clustered patterns (e.g., Cele CIRc 25, Atha CIRc 0 - see further) (Figure 5.4). Often, P interactions between the TFs are also present, as already indicated by the FBU/FFL/CIR motif overlap. In addition to feed forward loop modules, circular feedback modules form the core of integrated GRNs and integrate signalling between regulators, often coordinating developmental transitions.

In *C. elegans*, we also observed intrinsically clustered patterns of feed forward loops as well as circular feedback motifs, likely due to the higher clustering coefficients in the *C. elegans* data (Supplementary Data).

The sixth type, the feedback 2 undirected module (FB2Uc), is formed by clusters of protein-interaction mediated transcriptional regulatory loops, a motif that mediates undirected feedback between a TF and its target, through a common partner in the protein interaction network [78]. A first cluster generalization of this motif is the regulonic star, where multiple targets of a regulator interact with the same feedback protein, including the regulator itself [330]. In Cele FB2Uc 13 for instance, the transcriptional co-activator EYA-1 functions as feedback mediator in the development of various tissues [356]. A second cluster generalization is the regulatory interacting double-star, where one or a few regulator-target pairs share a common set of partners in the protein interaction network [330]. For example, in Cele FB2Uc 39, DAF-3 is a central protein interaction partner for a number of proteins, as well as a transcriptional regulator to DAF-7 and DAF-8 that both in return genetically interact with the protein interaction partners of DAF-3 (Figure 5.4). This extreme example combines these two types in a protein complex, where some members transcriptionally regulate other members (e.g. Atha FB2Uc 1 (Supplementary Data), Atha CIRc 0).

The seventh type, the feedback undirected module (FBUc), is similar to the feedback 2 undirected module, but now contains clusters of motifs consisting of two coherent regulatory edges and one undirected interacting edge. Here, we also detected the regulonic star, where now the feedback protein is another

regulator that targets the first regulator. In Atha FBUc 2, for instance, PIL5 and PIF4 target back to SEP3. Hence, a member of a protein complex also actively regulates its regulating TF. Due to the FBU/FFL and FBU/COR motif overlap, this module is also FFLc and CORc to some extent. Analogously, the regulatory interacting double star now consist of a regulator with a set of protein interaction partners that targets another regulator that transcriptionally regulates those protein interaction partners. In ATHA FBUc 3, MYB33 physically interacts with all the targets of LFY3 that is a direct target itself of MYB33 (Figure 5.4).

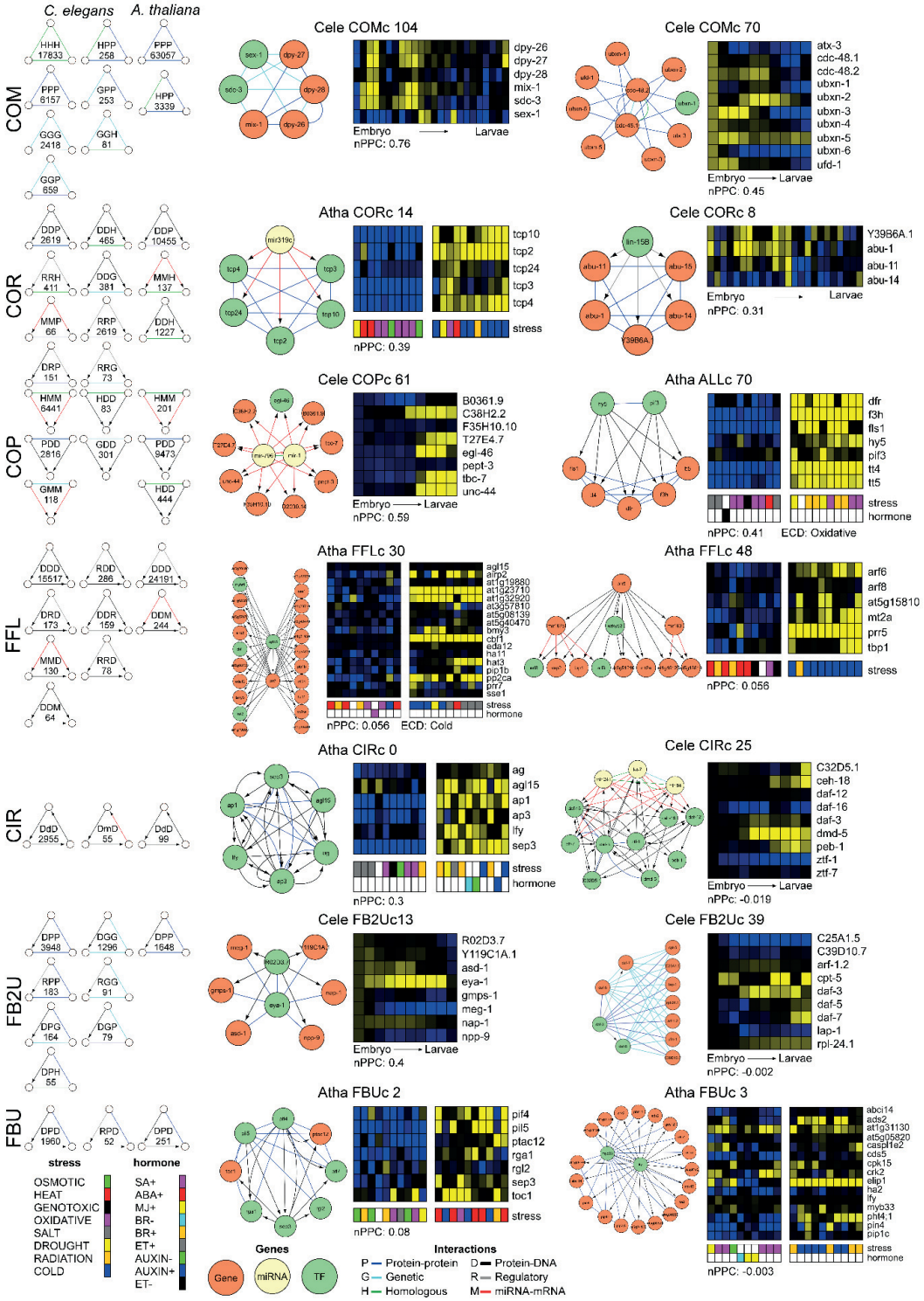

C. elegans
A. thaliana
COM
HHH 17833
HPP 258
PPP 63057
PPP 6157
GPP 253
HPP 3339
GGG 2418
GGH 81
GGP 659
COR
DDP 2619
DDH 465
DDP 10455
RRH 411
DDG 381
MMH 137
MMP 66
RRP 2619
DDH 1227
DRP 151
RRG 73
HMM 6441
HDD 83
HMM 201
PDD 2816
GDD 301
PDD 9473
GMM 118
HDD 444
COP
DDD 16517
RDD 286
DDD 24191
DRD 173
DDR 159
DDM 244
MMD 130
RRD 78
DDM 64
FFL
DdD 2955
DmD 55
DdD 99
CIR
DPP 3948
DGG 1296
DPP 1648
RPP 183
RGG 91
DPG 164
DGP 79
DPH 55
FB2U
DPD 1960
RPD 52
DPD 251
FBU
stress
hormone
OSMOTIC
HEAT
GENOTOXIC
OXIDATIVE
SALT
DROUGHT
RADIATION
COLD
SA+
ABA+
MJ+
BR-
BR+
ET+
AUXIN-
AUXIN+
ET-
Cele COMc 104
sex-1
dpy-27
sdc-3
dpy-28
mix-1
dpy-26
dpy-26
dpy-27
dpy-28
mix-1
sdc-3
sex-1
Embryo
Larvae
nPPC: 0.76
Cele COMc 70
ubxn-4
ufd-1
ubxn-2
ubxn-3
ubxn-1
cdc-48.2
ubxn-6
cdc-48.1
atx-3
ubxn-5
atx-3
cdc-48.1
cdc-48.2
ubxn-1
ubxn-2
ubxn-3
ubxn-4
ubxn-5
ubxn-6
ufd-1
Embryo
Larvae
nPPC: 0.45
Atha CORc 14
mir319c
tcp4
tcp3
tcp24
tcp10
tcp2
tcp10
tcp2
tcp24
tcp3
tcp4
stress
nPPC: 0.39
Cele CORc 8
lin-15B
abu-11
abu-15
abu-1
abu-14
Y39B6A
Y39B6A.1
abu-1
abu-11
abu-14
Embryo
Larvae
nPPC: 0.31
Cele COPc 61
egl-46
B0361.9
T27E4
mix-1
tbc-7
pept-3
unc-44
C38H2.2
F35H10.10
T27E4.7
egl-46
pept-3
tbc-7
unc-44
Embryo
Larvae
nPPC: 0.59
Atha ALLc 70
hy5
pif3
tt5
f3h
dfr
stress
hormone
nPPC: 0.41
ECD: Oxidative
Atha FFLc 30
agl15
alp2
at1g19880
at1g23710
at1g32920
at3g57810
at5g08139
at5g40470
bmy3
cbf1
eda12
ha11
hat3
pip1b
pp2ca
prr7
sse1
stress
hormone
nPPC: 0.056
ECD: Cold
Atha FFLc 48
arf6
arf8
at5g15810
mt2a
prr5
tbp1
stress
nPPC: 0.056
Atha CIRc 0
agl15
ap1
ap3
lfy
sep3
stress
hormone
nPPC: 0.3
Cele CIRc 25
C32D5.1
ceh-18
daf-12
daf-16
daf-3
dmd-5
peb-1
ztf-1
ztf-7
Embryo
Larvae
nPPc: -0.019
Cele FB2Uc13
R02D3.7
Y119C1A.1
asd-1
eya-1
gmps-1
meg-1
nap-1
npp-9
Embryo
Larvae
nPPC: 0.4
Cele FB2Uc 39
C25A1.5
C39D10.7
arf-1.2
cpt-5
daf-3
daf-5
daf-7
lap-1
rpl-24.1
Embryo
Larvae
nPPC: -0.002
Atha FBUc 2
pif4
pil5
ptac12
rga1
rgl2
sep3
toc1
stress
nPPC: 0.08
Atha FBUc 3
abi1
ads2
at1g31130
at5g05820
casp1e2
cds5
cpk15
crk2
elip1
ha2
lfy
myb33
pht4;1
pin4
pip1c
stress
hormone
nPPC: -0.003
Genes
Gene
miRNA
TF
Interactions
P Protein-protein
G Genetic
H Homologous
D Protein-DNA
R Regulatory
M miRNA-mRNA

Figure 5.4: ***Overview of the different network motif types (left) and modules for each motif type clustered (right)****. The number of specific motifs that were found at least 50 times in the GRNs of C. elegans and A. thaliana is indicated per motif type on the left. Specific examples of the clustering of motifs per motif type in modules is depicted on the right by a network figure and a Module Viewer figure of their expression profiles in developmental (C. elegans) or abiotic stress conditions (A. thaliana) (Methods). For the abiotic stress compendium Module Viewer figure, only the top 10 conditions with most up- and down-regulated expression are shown. The average Pearson Correlation Coefficient (nPCC) and if available, the abiotic stress condition with significant Expression Correlation Differential score (ECD) are shown as measures of coexpression and expression dynamicity of the modules, respectively.* ***Complex motifs (COM) and modules (COMc):*** *Cele COMc 104 (GGG/GPP/PPP/GGP motifs) involved in dosage compensation and sex determination and Cele COMc 70 (PPP/GPP/HPP motifs) functioning in ubiquitin-dependent protein catabolism.* ***Co-regulated motifs (COR) and modules (CORc):*** *Atha CORc 14 (MMP/PPP motifs) involved in leaf and flower development [357], upregulated upon cold stress and downregulated upon oxidative stress, and Cele CORc 8 (RPP/PPP motifs) involved in the endoplasmic reticulum unfolded protein response.* ***Co-pointing motifs (COP) and modules (COPc):*** *Cele COPc 61 (HMM motifs) involved in axon extension and Atha ALLc 70 (DDP/PDD/PPP motifs) involved in flavonoid biosynthesis, upregulated upon radiation stress and dynamic upon oxidative stress.* ***Feed-forward motifs (FFL) and modules (FFLc):*** *Atha FFLc 30 (DDD motifs) involved in response to water deprivation, upregulated upon cold and salt stress and dynamic upon cold stress, and Atha FFLc 48 (DDD/DMD motifs) upregulated upon cold and salt stress, downregulated upon heat stress.* ***Circular feedback motifs (CIR) and modules (CIRc):*** *Atha CIRc 0 (DdD/DDD/DPD/PPP/DPP motifs) involved in flower development and Cele CIRc 25 (DmD/DdD/DDD/DPP/DPD/RPD/RDD/DMM/DDG/DDH/DDP/DMD/GHH/HMM/GMM motifs) involved in the regulation of larval development.* ***Feed-back 2 undirected motifs (FB2U) and modules (FB2Uc):*** *Cele FB2Uc 13 (DPP motifs) involved in embryonic and larval development and Cele FB2Uc 39 (RPG /RGP/GGG/GPP motifs) involved in Dauer larval development.* ***Feed-back undirected motifs (FBU) and modules (FBUc):*** *Atha FBUc 2 (DPD/DDD/PPP/PDD/DDP motifs) involved in the cellular response to red or far red light and upregulated upon heat and cold stress and Atha FBUc 3 (DPD motifs) functioning in flower development.*

In the ALL modules, all motif types were clustered together. Here, we typically found integration of different motifs and combinations of different modules. As an example, in a merged module, which consists of COP, COR, and FFL motifs, and functions in flower development, the homologous miRNAs miR156/157 post-transcriptionally regulate members of the squamosa-promoter binding protein-like (SPL) gene family (Figure 5.5A, Suppl. Fig. 6). This miR156/157-SPL module is a regulatory hub important for the transition from vegetative phase into flowering. It is closely linked with environmental signals like temperature, salt and light [358]. On top of that, SEPALLATA3 (SEP3) targets both miRNAs and SPLs, creating TF-mediated miRNA FFLs. From literature, it is known that SEP3 is a responsive gene of SPL3 in the ambient temperature-responsive flowering [359]. Here, we observed that SEP3 also functions as an upstream regulator by binding other SPL TFs that are upregulated upon abiotic stress.

In the partially overlapping modules of Atha ALLc 93, ALLc 147, COMc 14, COPc 11, COPc 47 and COPc 14, which combine COP, COR, COM, and FFL motifs, a gene family of eleven zinc finger homeodomains interact through P and H interactions, while several of them are targeted by the flowering regulator AGL15 and/or transcriptionally regulate genes involved in secondary cell wall and glucosinolate biosynthesis (Figure 5.5B). In Arabidopsis, zinc finger homeodomains are known to homo- and heterodimerize and play overlapping regulatory roles in floral development [360]. We found that the HB-genes in the protein cluster COMc 14, COPc 14 and ALLc 93 are significantly co-expressed (Suppl. Fig. 7). We also observed that the highly similar homologs hb30/hb34 are expressed under the same abiotic stress conditions and that they are dynamically expressed in osmotic stress conditions in roots together with zfhd1 of which the upregulation by high salinity was already reported [361] (Figure 5.5B, Suppl. Fig. 7). Furthermore, we perceived in the abiotic stress expression compendium different expression preferences for the different zinc finger homeodomains upon osmotic, salt, heat and cold stress: most of them are preferentially expressed in root tissues, only hb23 also

shows expression in shoot tissues (Figure 5.5B). Together this leads to the assumption that they, despite the large functional overlap between the hb-genes, have diverged to regulate development under specific abiotic stresses. Differential expression of zinc finger homeodomains under abiotic stress conditions has already been shown in *Brassica rapa* and *Vitis vinifera* [362, 363]. Therefore, it is interesting to study the evolutionary diversification of these zinc finger homeodomains in the abiotic stress response.

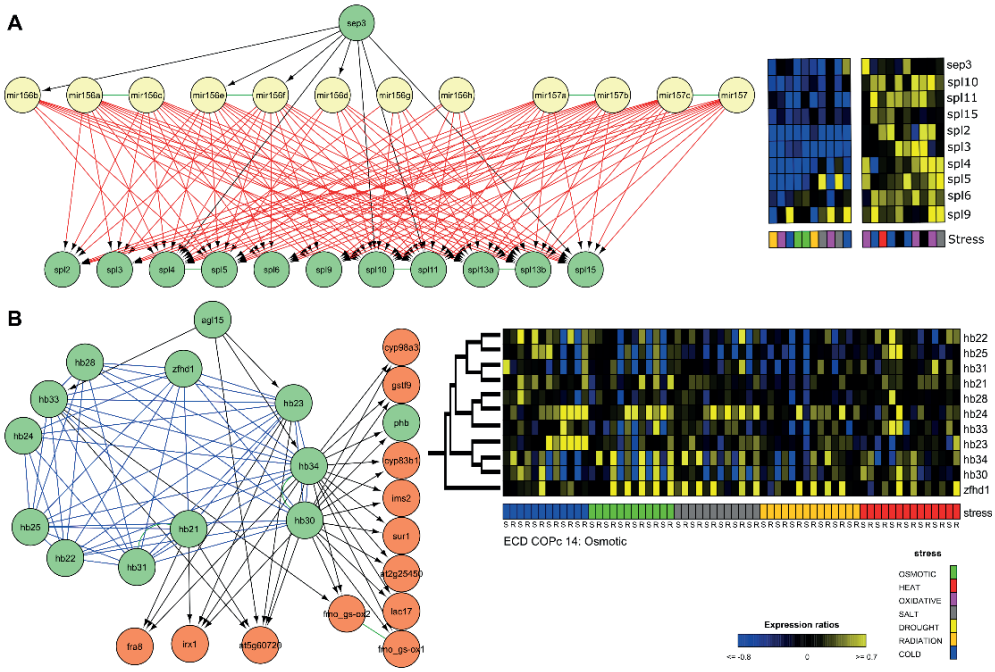


Figure 5.5: ***ALL modules formed by clustering different motif types together****. Expression profiles are depicted for the 10 abiotic stress conditions with maximal up- and downregulation.* ***A)*** *Merged FFLc, COPc and CORc module from ALLc 50 largely overlapping with COPc 10 and CORc 5. Here, the homologous miRNAs miR156/157 post-transcriptionally regulate members of the squamosa-promoter binding protein-like (SPL) gene family. Both miRNAs and SPLs are targeted by SEP3. The functional diversification of the different spl-genes is illustrated by the mixture of different stresses in the conditions with maximal up- and downregulation for the abiotic stress compendium.* ***B)*** *Overview of the different modules with zinc finger homeodomains: ALLc 147, COMc 14, COPc 11, COPc 47 and COPc 14. Experiments with specific tissues, root (R) and shoot (S), are marked below the expression profile matrix. Genes are sorted based on the gene family tree [360].*

The advantage of our data-integration methodology, which captures different experimental methodologies and resources, is, for example, shown in the integrated complex modules of Arabidopsis SWI/SNF chromatin remodelling complexes (Suppl. Fig. 8) and the *C. elegans* coregulated module Cele CORc 26 (Suppl. Fig. 9). The SWI/SNF chromatin remodelling modules are formed by complexes that interact with each other and consist of protein-protein interactions gathered by Y2H, tandem-affinity purification (TAP), protein-fragment complementation assay (PCA) and other techniques [219, 364, 365]. According to their experimental methodology, TAP detected the SWI/SNF complex around the central ATPases BRM or SYD [219], while Y2H identified binary interactions between the SWI/SNF subunits and several TFs and cofactors (Suppl. Fig. 8). In *C. elegans* coregulated module Cele CORc 26 with data-integration of ChIP and Y1H, Y1H has the highest incoming degree, while ChIP has the highest outgoing degree in the module, as can be expected from their experimental methodology (Suppl. Fig. 9).

Overall, we found similar network motif modules in the integrated GRNs of *C. elegans* and *A. thaliana*, suggesting these topological patterns are universal in networks of gene regulation. A dynamic visualization of all modules can be found on http://bioinformatics.psb.ugent.be/supplementary_data/jofoo/networks/. This interactive visualization groups all modules per type with links to the expression matrices. Through the search bar it is possible to look for genes of interest in both species.

5.4.4 A superview analysis of network motif modules

The network motif modules are part of integrated GRNs, where they influence one another and might be active under different conditions. We developed a method to investigate modules in the network context, where we studied interactions between the modules and regulators through statistical analysis to find enrichment for functional and regulatory important edges (Methods) (Figure 5.1). Linking the modules through homologous interactions and/or shared genes results in groups of modules involved in similar processes. For example, in *A. thaliana* we found six alternative splicing modules connected through homology edges and controlled by abiotic stress (Suppl. Fig. 10). In addition, we looked for TF and miRNA regulators specifically targeting one or more modules (Suppl. Fig. 11). In a first example we confirmed the regulation of the cellulose synthase complex (CSC) COMc 36 by MYB46 in *A. thaliana* [366] (Figure 5.6A).

In a second example, we illustrated that the superview framework is able to highlight unexplored module-regulator connections. Here, the homeodomain TF CEH-30, which functions in neuronal cell fate and sex-specific apoptosis, targets a homolog group of heat shock proteins in worm (Cele COMc 35) (Figure 5.6B). Finally, we also found novel targets for known regulators. We link CBF4, a regulator of the ABA dependent drought response [367], and ZML2, a critical TF in the cry1-mediated photoprotective response [368], to aliphatic and indolic glucosinolate biosynthesis in Atha COMc 48 (Figure 5.6C). This module has a significant ECD in drought and salt stress.

These examples show how the network motif modules can be integrated into a larger context beyond individual modules and how general topological patterns can enable the study of stress related mechanisms. Through usage of different expression compendia or additional regulatory data, other processes can be explored as well.

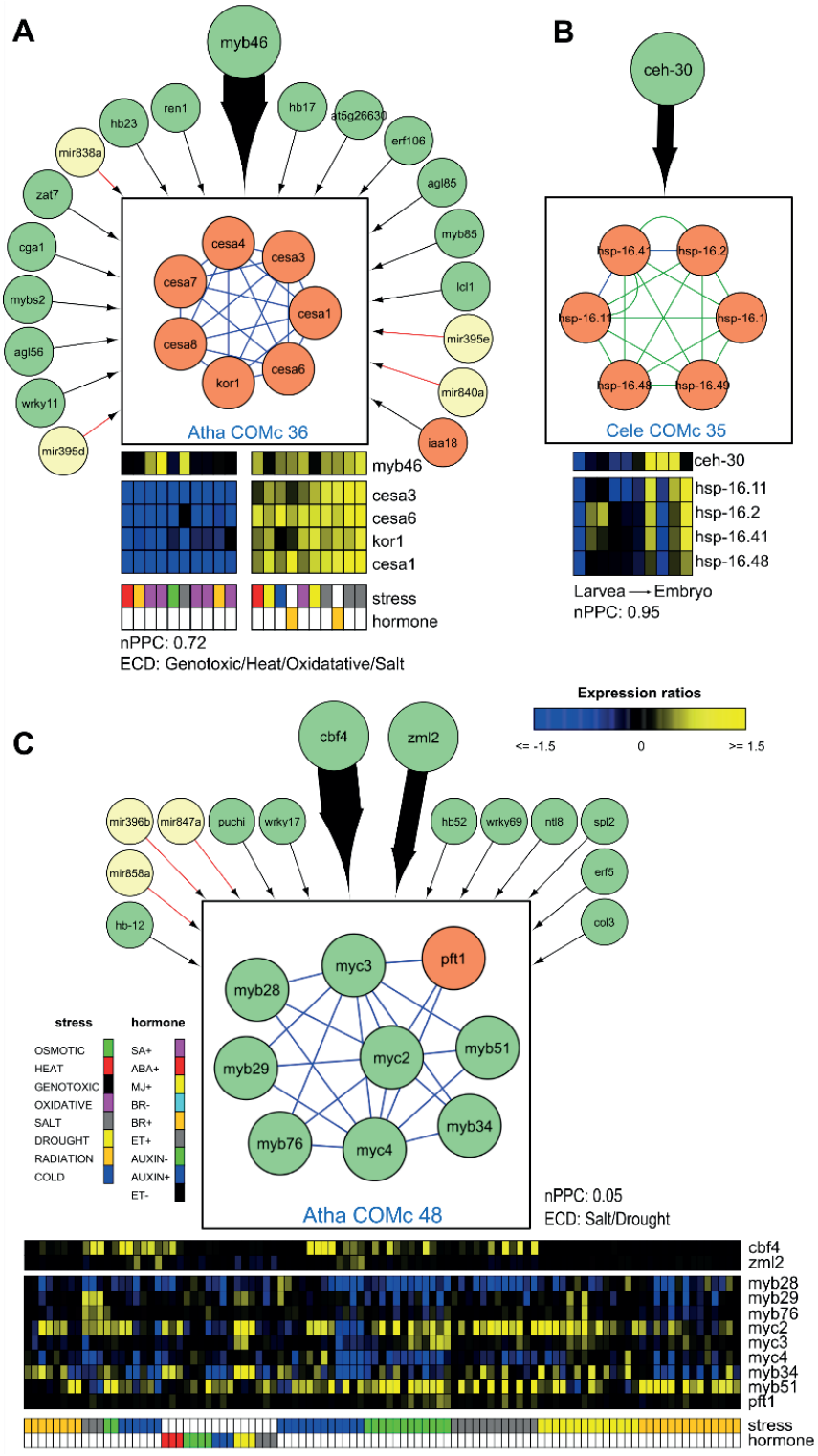

A
myb46
hb23
ren1
hb17
at5g26630
erf106
mir838a
agl85
zat7
myb85
cga1
lcl1
mybs2
mir395e
agl56
mir840a
wrky11
iaa18
mir395d
cesa4
cesa3
cesa7
cesa1
cesa8
cesa6
kor1
Atha COMc 36
myb46
cesa3
cesa6
kor1
cesa1
stress
hormone
nPPC: 0.72
ECD: Genotoxic/Heat/Oxidatative/Salt
B
ceh-30
hsp-16.41
hsp-16.2
hsp-16.11
hsp-16.1
hsp-16.48
hsp-16.49
Cele COMc 35
ceh-30
hsp-16.11
hsp-16.2
hsp-16.41
hsp-16.48
Larvea ⟶ Embryo
nPPC: 0.95
C
Expression ratios
<= -1.5
0
>= 1.5
cbf4
zml2
mir396b
mir847a
puchi
wrky17
hb52
wrky69
ntl8
spl2
mir858a
erf5
hb-12
col3
myc3
pft1
myb28
myb51
myc2
myb29
myb34
myb76
myc4
Atha COMc 48
stress
hormone
OSMOTIC
HEAT
GENOTOXIC
OXIDATIVE
SALT
DROUGHT
RADIATION
COLD
SA+
ABA+
MJ+
BR-
BR+
ET+
AUXIN-
AUXIN+
ET-
nPPC: 0.05
ECD: Salt/Drought
cbf4
zml2
myb28
myb29
myb76
myc2
myc3
myc4
myb34
myb51
pft1
stress
hormone

Figure 5.6: ***Through the superview analysis framework, we discovered previously known (A) and unknown (B) regulators for specific modules, as well as (C) novel edges for known regulators****.* ***A)*** *Cellulose synthase complexes (CSC) in COMc 36 are upregulated by MYB46. While MYB46 binds 4 module genes, the other regulators bind only one gene in the module. The module consists out of the primary cell wall CSC (CESA3, CESA1 and CESA6), the secondary cell wall CSC (CESA4, CESA7, and CESA8) and KOR1, a membrane-bound 1,4-beta-D-glucanase [369, 370]. This module is tightly co-expressed in the abiotic stress compendium and upregulated upon brassinosteroid treatment [371] and salt stress conditions. COMc 36 has a significant ECD score under genotoxic, heat, oxidative, and salt stress. In birch, overexpression mutants of MYB46 show thicker secondary cell walls and a higher tolerance to salt and osmotic stress [372]. Cellulose synthases bind microtubules, hence stabilizing cellulose synthase localization at the plasma membrane and rendering plants less sensitive to salt stress [373]. The relation between MYB46 and CSC is therefore important for the stress tolerance of crops. This example highlights the potential of integrating regulators with network motif modules.* ***B)*** *The homeodomain TF CEH-30, which functions in neuronal cell fate and sex-specific apoptosis, was found to target a homolog group of heat shock proteins in worm.* ***C)*** *CBF4 and ZML2 transcriptionally regulated the MYB/MYC module Atha COMc 48. The TFs MYB28, MYB29 and MYB76 control aliphatic glucosinolate biosynthesis [374], while MYB51 and MYB34 regulate indole glucosinolate biosynthesis [375]. The JAZ-interacting TFs MYC2, MYC3 and MYC4 form together with the MYB TFs dimeric TF complexes to regulate the different glucosinolate biosynthesis pathways [376]. Glucosinolates, a class of secondary metabolites mainly found in Brassicaceae, are part of a complex response to a variety of abiotic stresses. A decrease in aliphatic glucosinolates modifies the abundance of aquaporins and hence the water uptake in roots, thereby increasing drought and salt tolerance [377]. Only the aliphatic glucosinolate biosynthesis TFs are directly bound by CBF4. In our abiotic stress compendium, we observed an upregulation of aliphatic glucosinolate biosynthesis (MYB26 & MYB76), indolic glucosinolate biosynthesis (MYB51), MYC2, and also of CBF4 upon salt stress; for MYB51 and CBF4 this is mostly in roots. It has been observed that CBF4 significantly alters the accumulation of at least five glucosinolates but the direct regulatory mechanism between CBF4 and glucosinolate synthesis has not been described [378]. Here we showed that the drought responsive gene cbf4 is an upstream regulator of the aliphatic glucosinolate biosynthesis which increases the tolerance to drought and salt stress. The function of zml2 in this context is still to be determined.*

5.4.5 Phylogenetic decomposition of the networks

Through phylogenetic decomposition of these integrated GRNs, we investigated how novel genes are integrated in GRNs. Therefore, genes were arranged in age groups or phylostrata based on the oldest lineage that still contained an ortholog (Suppl. Fig. 12, Methods). This resulted in respectively 7 and 10 age groups for *C. elegans* and *A. thaliana* (Suppl. Tables 10 and 11). 61% of *C. elegans* and 99% of *A. thaliana* protein-coding genes in the GRNs could be given an age label. In the worm integrated GRNs, the groups Eukaryota and Caenorhabditis each contain more than 25% of all age-labelled genes, the groups Eumetazoa and Cellular organisms each have around 15% of these genes, while the other age groups each take 5% or less. In the plant integrated GRNs, nearly half of all age-labelled genes reside in the oldest age group Green plants, followed by 27% in Land plants, 7% each in Seed and Flowering plants and less than 5% in the other age groups. We restricted ourselves to P, G, D and R interactions in the networks: 44% and 99% of *C. elegans* and *A. thaliana* interactions respectively, have associated age labels. For worm, the age groups with most interactions were Eumetazoa (36%), Eukaryota (30%) and Caenorhabditis (15%) (Figure 5.7). For Arabidopsis, interactions are concentrated in Green plants (46%), Land plants (30%) and Flowering plants (10%) (Figure 5.7, Suppl. Table 12). Hence, the interactions are mainly distributed over the age groups containing the most genes. Among these age groups are the oldest ones like Eukaryota in worm and Green and Land plants in plant. Another reason for the interaction distribution is the fact that older genes are better studied than young genes and therefore more represented in the networks for both species (Supplementary Data, Suppl. Tables 10 and 11, [337, 338]). The average degree is mostly confirming these observations (Suppl. Table 13). For worm, the highest average undirected, incoming and outgoing degree are observed for the Eumetazoa. The further away from this age group, the lower the degrees become. For plant, the highest average undirected degrees are seen in the Land and Flowering plants, although with the exception of Rosids and *A.*

thaliana, other age groups have only slightly lower average undirected degrees. Although average incoming degrees are similar for all plant age groups, the average outgoing degree of the Flowering plants towers.

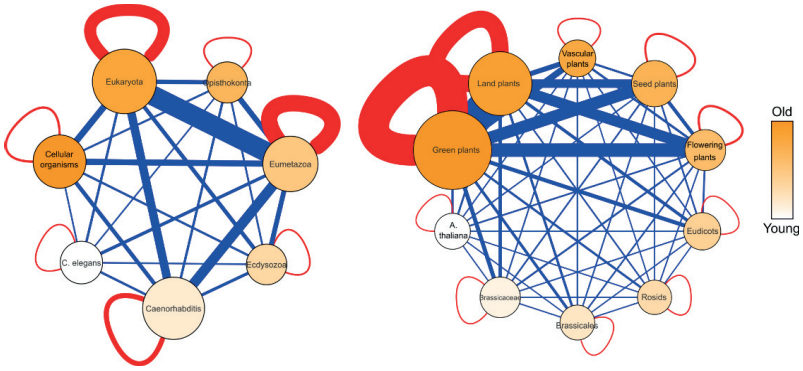


Figure 5.7: ***Total number of directed and undirected interactions between age groups of A. thaliana (left) and C. elegans (right).*** *The nodes are scaled according to the number of genes in the age group and coloured according to age (darker = older). Red edges are within the age groups, blue edges are between the age groups. The thickness of the edge is scaled to the number of interactions, which is also mentioned in the edge label for the interactions within age groups.*

5.4.6 Protein-protein interactions preferentially occur between proteins of similar age, while for protein-DNA interactions, regulatory TFs favour older or same-age target genes

To investigate the general interaction preference of the different types of molecular interactions in our integrated GRNs, for each interaction type we analysed whether they preferred to interact within or between age groups. In both species, we found that P interactions are preferentially age homogeneous. In *A. thaliana* and *C. elegans*, respectively 40% and 34% of the interaction partners are of the same age. This is significantly more than expected by random (Atha $p < 2.2e\text{-}16$; Cele $p < 2.2e\text{-}16$, Z-test, 1000 random network permutations with preserved age and degree distribution). The interaction partners of protein-DNA interactions were less frequently of the same age, but still significantly more than expected by random (Atha age homogeneous only ChIP and Y1H: 31%, p-value = 0.0034; Cele age homogeneous D: 30% p-value = 0.036, Z-test). The full D set of *A. thaliana*, which includes the regulatory interactions, did not prefer to interact with genes of similar evolutionary age (Atha age homogeneous 25.2%, p-value = 0.50, Z-test), as well as the regulatory (Cele age homogeneous: 25.8% p-value = 0.36, Z-test) and genetic interactions (Cele age homogeneous: 33.4% p-value = 0.073, Z-test) in *C. elegans*. This is understandable, since the latter interactions are not necessarily direct interactions that might involve intermediate nodes in the networks.

To investigate the interaction preference in relation to evolutionary age in the integrated GRNs, for both species we compared the number of observed interactions within and pairwise between the different age groups versus the expected number of interactions based on randomized networks with the same age distribution as the real networks. Due to the differences in age homogeneity for the different interaction types (see above), we here show the results of the physical protein-protein and protein-DNA interactions, while the results of the whole set of undirected and directed interactions can be found in the Supplementary Data. This analysis indicated that some age groups attracted many more interactions than expected by random. In both *C. elegans* and *A. thaliana* protein-protein interaction networks, we observed an interaction

age preference towards the own age group or to the next age groups i.e. the highest Z-scores are found on or near the main diagonal of the age group matrix (Figure 5.8). These results are confirmed for the undirected networks data (Suppl. Fig. 13) and through the calculation of the interaction density (Supplementary Data). Hence, proteins prefer to functionally interact with proteins of similar evolutionary age. Since the overlap between the homologous interactions and the other type of interactions is small (Suppl. Fig. 1), we can exclude that the interaction preference of genes in the same age group originates from interactions between homologs.

Considering the protein-DNA interaction networks in worm, we noted strong preferences of Ecdysozoan TFs for target genes from Cellular Organisms and Eukaryota, of Eukaryotic TFs for Eukaryotic target genes, and of Caenorhabditis TFs for Eumetazoan target genes (Figure 5.8). For plant physical protein-DNA data, regulatory TFs from Eudicots or older age groups preferred to bind target genes from the Green or Land plants (Figure 5.8). We found similar results for the directed networks (Suppl. Fig. 13) and through the interaction density analysis (Supplementary Data). Hence, for the protein-DNA interaction networks in worm and plant, the highest Z-scores are found on or above the main diagonal in the age group regulatory TF-target gene matrix, indicating that regulatory TFs tend to bind target genes of similar or older evolutionary age.

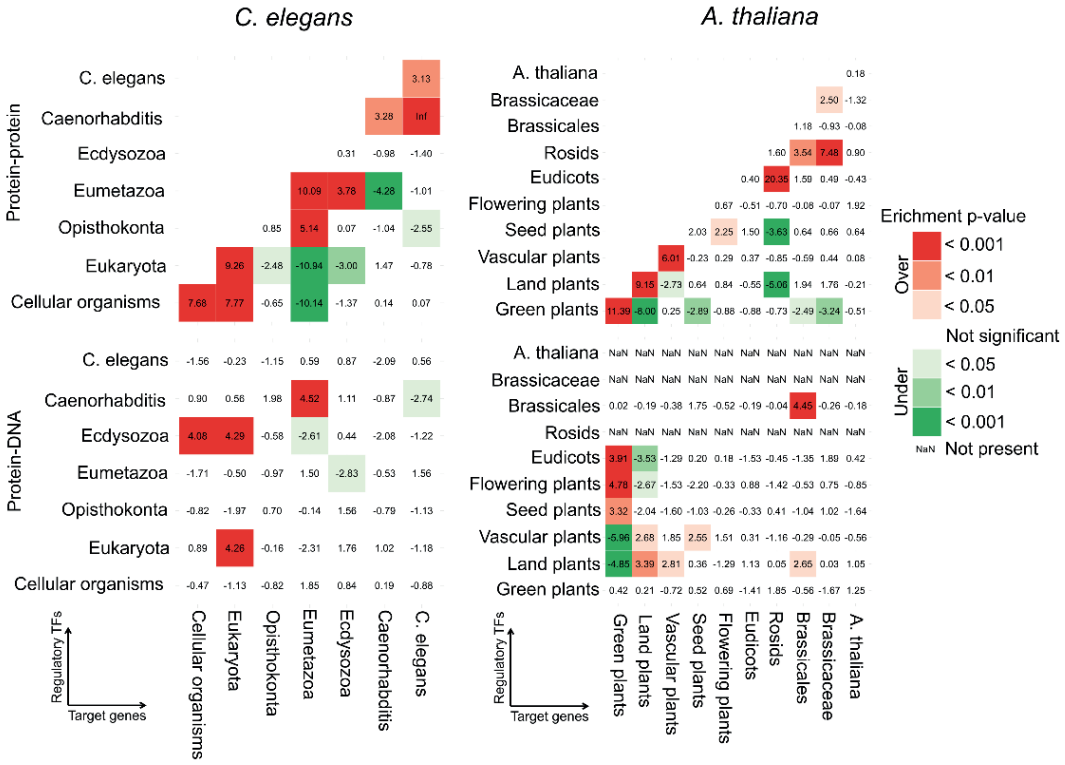


Figure 5.8: ***Interaction age preference for physical protein-protein and protein-DNA interactions in C. elegans (left) and A. thaliana (right)****. Enrichment (Z-score and associated p-value corrected for multiple hypothesis testing) of the comparison of the observed number of interactions within and between age groups in the real networks versus the expected number in 1000 randomized networks with the same age distribution. In the protein-DNA networks, the regulatory TFs are on the vertical and the target genes are on the horizontal axis.*

5.4.7 Interaction age preference of motifs and modules

Since motifs are considered to be small building blocks of networks, we investigated how novel genes are incorporated in integrated GRNs at the motif level. Therefore, motifs were divided into 13 motif age types based on the age pattern of the different node positions: either all nodes are of the same phylostratum (SSS, S=Same), there are two age groups in the motif (Y=Young and O=Old) or they are all from a different phylostratum (Y=Young, M=Middle and O=Old). Motifs with internal symmetry were sorted according to decreasing age to remove overlap between the motif age types. We calculated motif age preference by comparison of the observed motif age patterns to the expected patterns by permuting each of the three node positions with preserved age distribution per node (Methods, Suppl. Table 16). For the complex motifs, we observed a strong preference to be age homogeneous in both species, especially in *A. thaliana* (Figure 5.9A). Also, in *C. elegans*, at least one edge in the complex motifs is between proteins of similar age. Similarly, the co-regulated motifs (COR) tended to be completely age homogeneous (SSS-type) or of the OYY/YOO-type, where the targeted nodes are of similar age and interact undirectedly. In addition to the preferentially age homogeneous type (SSS-type), both the plant and worm co-pointing motifs (COP) were composed of the YYO age type, where two younger TFs of the same age interact and target a gene of a different phylostratum. In addition, we observed a strong enrichment for the OYM age type in plant COP motifs, where a physical bound between an old and young TF regulates a middle-aged target gene. The feed-forward loops (FFL) showed the strongest preference to be age heterogeneous: OYM was the strongest age motif type enriched in both species, followed by YOO and MOY in plant, and by SSS, MOY and YYO in worm. Hence, novel genes are incorporated at every position in the FFL. CIR motifs were preferential age homogeneous in *C. elegans* or of the heterogeneous OMY-type in both species. The FB2U motifs followed mostly the same trends as the complex motifs while the FBU motifs were similar to the FFL motifs. This can be explained by the overlap between these motif types (Suppl. Fig. 5). In *C. elegans*, almost all motifs displayed enrichment in the homogeneous motif age type due to the dominance of the Eumetazoa interactions, e.g. the DDD motif consists out of 23% Eumetazoa SSS type and only 0.67% other SSS-type motifs. Overall, we observed that undirected interactions in motifs tended to be age homogeneous, while directed interactions in motifs preferred to be age heterogeneous.

Several of the observed motif age types can originate from gene duplication. To investigate the contribution of duplicates to motif formation, we first looked at the number of motifs consisting out of at least one pair of homologs. We observed that homologous genes only appeared in at most 2% and 1% of motifs excluding H interactions in *C. elegans* and *A. thaliana*, respectively (Suppl. Table 17 and Suppl. Table 18). They appeared together in up to 6% of all DdD, DDD, DDP, DPD, DPP, PDD and PPP motifs in both species. Secondly, we compared the number of genes with H interactions in the complete interaction set versus in the motifs (Suppl. Table 19). For protein-protein, regulatory and genetic interactions, we found no preferential motif formation of genes with homologous interactions in both species. For protein-DNA interactions, we found that genes with homologous interactions contribute more to motif formation in *A. thaliana* than expected.

Different network motifs have specific evolutionary age types associated. To investigate whether the preferred age patterns in the motifs are also preferentially incorporated into the modules upon motif clustering, we compared the set of clustered motifs to the full set of motifs (Figure 5.9C). In both species, we found in the COMc, COPc, CORc, and FFLc modules a strong correspondence between the overrepresented

motif age types in their underlying network motifs and those that are clustered in their modules. In the CIRc and *A. thaliana* FBUc modules, we observed no real preference of clustering because almost all motifs are clustered within the modules. In FB2Uc and *C. elegans* FBUc modules, age heterogeneous motif types are preferentially incorporated. The age homogeneous motif types SSS in *C. elegans* are less clustered than expected in almost all modules. This might be explained by the overrepresentation of age homogeneous motif types from the Eukaryota and Eumetazoa. Overall, we observed that the overrepresented age motif types clustered more than the other types in the modules i.e. we observed similar patterns in Figure 5.9C as compared to Figure 5.9B.

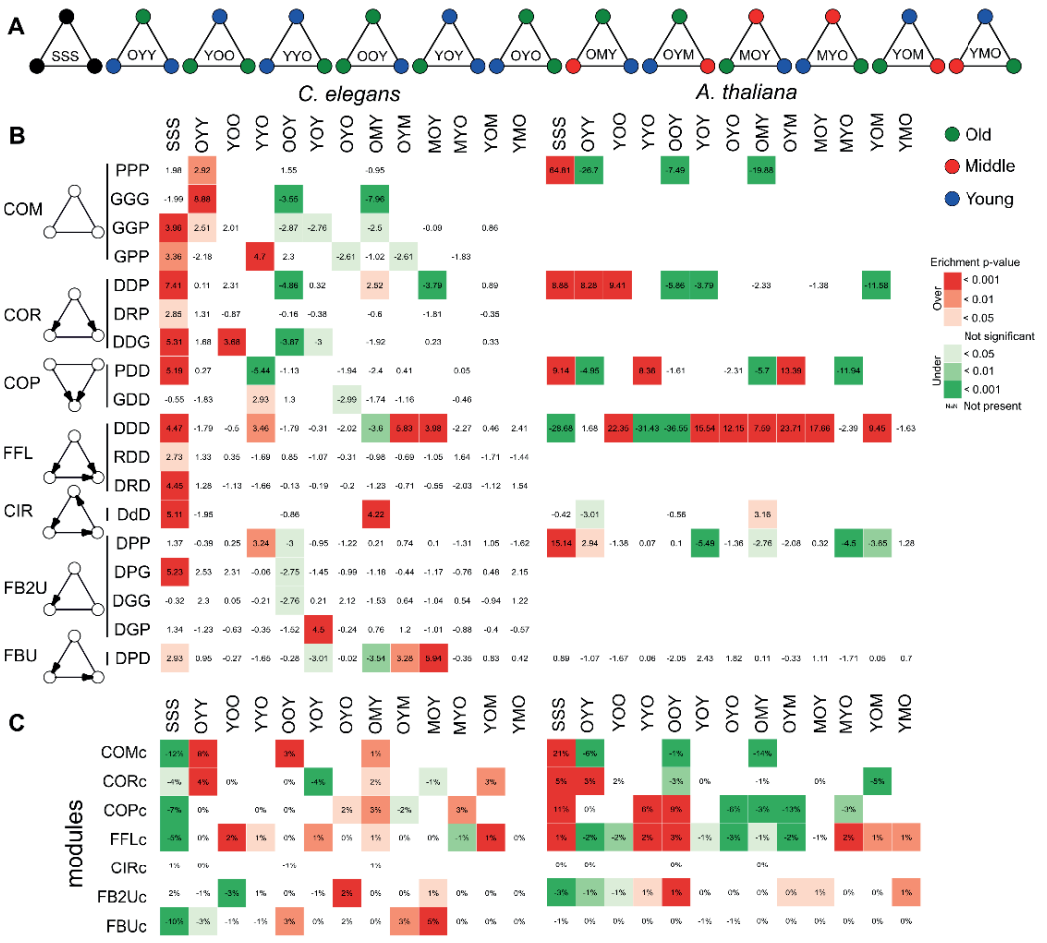


Figure 5.9: ***A)*** *The different motif age types* ***B) Motif age preference****. Statistical significance (empirical Z-score) of the observed age patterns in motifs compared to the patterns expected by random. Due to symmetry, not every pattern is present in all motifs (blank squares). The symmetric motif age types where sorted form old to young age. Only motifs with at least one significant after multiple hypothesis correction (Benjamini & Hochberg, p value < 0.05) observation are shown in the picture, the full Table can be found in Suppl. Table 15.* ***C) Module age preference.*** *Preferential age patterns of the motifs clustered in network motif modules. The value represents the percentage of motifs with each age pattern that are clustered subtracted by the percentage of a certain age pattern in all the motifs belonging to that module type. The squares are coloured according to the significance of this value (hypergeometric test with multiple hypothesis correction according to Benjamini & Hochberg). Due to symmetry, not every pattern is present in all motifs (blank squares). The symmetric motif age types where sorted form old to young age.*

The evolutionary age groups contributed differently to the modules. Modules were mostly composed of genes from the evolutionary age groups containing the most genes i.e. Eukaryota, Eumetazoa and Caenorhabditis in worm, and Green and Land plants in plant (Suppl. Fig. 14AB). Hence, especially in *A. thaliana*, younger genes were more inclined to attach to modules mostly composed out of older genes instead of forming modules on their own. Looking at the individual module types we noted that there are COMc modules that are age homogeneous in the older groups, Green and Land plants in *A. thaliana* and Eukaryota and Eumetazoa in *C. elegans* (Suppl. Fig. 14CD). In the other modules, there is little contribution of the other younger age groups. Age homogeneous modules with directed interactions are less abundant and appeared within the oldest group of *A. thaliana* i.e. Green and Land plants and within the Eumetazoa in *C. elegans*. This is in agreement with the preferential clustering of more age heterogeneous motifs in these regulatory modules (Figure 5.9C).

Atha COMc 48, already discussed above, is a prime example of how innovation is introduced in gene regulatory networks (Fig 10). Indolic glucosinolate biosynthesis originated in the Land plants, and therefore the indolic glucosinolate biosynthesis TFs (MYB51 and MYB34) belong to the Land plants phylostratum. Together with the JAZ-interacting basic helix-loop-helix TFs MYC2, MYC3 and MYC4, they form heterodimer TFs that transcriptionally activate glucosinolate biosynthesis genes. From Brassicales on, not only indolic, but also aliphatic glucosinolates appeared as secondary metabolites [379]. Therefore, the aliphatic glucosinolate biosynthesis TFs (MYB28, MYB29, and MYB76), which belong to the Brassicales phylostratum, were introduced in the GRNs through interactions with the MYC TFs.

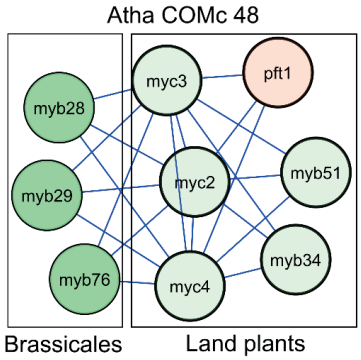


Figure 5.10: ***Atha COMc 48 coloured according to evolutionary age with older genes being more transparent****. (Functional interpretation and superview of module can be found in Figure 5.6C). The aliphatic glucosinolate biosynthesis TFs (MYB28, MYB29, and MYB76) find their origin in the Brassicales age group (R2D3-MYB subgroup 12), while the indolic glucosinolate biosynthesis TFs (MYB51 and MYB34) and MYC TFs (MYC2, MYC3 and MYC4) have their origin in the Land plants. They are split off from the rest of myb gene family which originate at the base of the Land plants. This is consistent with the observation that aliphatic glucosinolates are only found within the Brassicales plant lineage.*

5.5 DISCUSSION

5.5.1 Data integration through network motif modules

Since different molecular interaction types influence gene regulation, we developed a general data integration framework to study integrated GRNs of directed protein-DNA, transcription regulatory, miRNA-mRNA interactions and undirected protein-protein, genetic and homologous interactions. Our data integration framework of composite network motif modules is unbiased, since it does not favour any interaction type or experimental methodology over the other, and preserves the identity of the interaction type as compared to other data integration methodologies that benchmark using true positive data sets, Gene Ontology or KEGG [380-382]. The integration of complementary data types through 2- and 3-node motifs provides useful insights in the study of gene regulation and in GRN evolution. Motifs, like the well-described feed-forward loop, connect the regulatory levels (transcriptional and posttranscriptional) and integrate the directed and undirected interactions into easy interpretable patterns of gene regulation. Also, the incorporation of homologous interactions in motifs provides insights in how motifs and networks are formed by evolution. Next to the already integrated interactions, the network could still be expanded with epigenetic regulation and post-translational modifications, which are also known to affect gene regulation [322, 383].

Contrary to previous data integration studies [320], we also highlighted the effects of combining different experimental methodologies in the protein-protein interaction networks (Suppl. Fig. 8) and in the protein-DNA interaction networks (Suppl. Fig. 9, Suppl. Table 3 & Supplementary Data). One advantage is that different methodologies are complementary for a given data type and provide a more holistic view on gene regulation. For example, the integration of Y1H and ChIP data created extra CIR motifs (DdD, DmD) in the worm networks, indicating that there is possibly condition-dependent feedback regulation at the transcriptional and posttranscriptional level. Although we barely detected the 2-node miRNA-TF feedback loop in the networks, which is contrasting to other studies that used lower computational cut-offs [309], we found the 3-node miRNA-TF feedback DmD in the Y1H and in the combined Y1H and ChIP networks of worm. Hence, an intermediate regulatory TF confers the feedback of a TF to the miRNA it is regulated by. The joining together of different experimental methodologies also poses some challenges, as demonstrated by the biases introduced in network randomization, and hence network motif enrichment. The best-known motif in GRNs, the feed forward loop (FFL/DDD) [77], despite its abundance in both species and its important regulatory characteristics, is not found to be enriched in the integrated GRNs of both species, and only in the ChIP data of both species, as has been observed previously [320]. We also noted other differences in network motif enrichment between Y1H, ChIP and the combined data (Supplementary Data). We hypothesize that this different network motif enrichment can be mainly attributed to the edge swapping randomization of the networks, which has drawn criticism before [384-386]. Edge swapping randomization while preserving the degree distribution limits the randomization options for hubs and this affects the experimental methodologies differently. Since Y1H and ChIP data generate a different network topology with respectively 1-3 times more regulators than targets in Y1H and 5-20 times more targets than regulators in ChIP; more target gene versus regulator hubs, a higher clustering coefficient and a higher overall centrality for Y1H as compared to ChIP, this results in different randomized networks and therefore different network motif enrichment (see Supplementary Data). As network motif enrichment is highly sensitive to experimental

methodology, network topology and randomization, we recommend to study network motif presence and aggregation into modules.

Overall, we found the same 3-node motif types in both species. COM, COP, COR, FFL, and FB2U motifs were already detected previously [78, 314, 320, 329], additionally we detected the CIR motif where three regulators act upon each other through a feedback loop and the FBU motif where feedback to a linear path of directed edges is provided by an undirected interaction (Figure 5.4). In both species, both network motifs at the transcriptional level largely overlapped with the FFL DDD (Suppl. Fig. 5), indicating that intricate regulation between TFs occurs through feedback loops consisting of both transcription regulatory and physical protein-protein interactions. As we also incorporated miRNA-mRNA interactions, we also found the miRNA-mediated FFL (MMD) and the TF-mediated miRNA FFL (DDM) at the posttranscriptional level [309, 319-321].

Although network motifs are basic building blocks of GRNs, several studies have pointed out that aggregation of motifs into larger modules occurs naturally and might be more important to consider, not only from a topologically point of view, but also functionally and evolutionary [326, 327, 330, 331]. The module level is also claimed to be the most conserved one across species [387]. Therefore, our data integration framework focused on network motif modules. We were able to detect topological organizations of integrated GRNS which are similar in *C. elegans* and *A. thaliana*. The network motif modules, COMc, COPc, CORc, FFLc and FB2Uc have been described in yeast and were previously detected either based on visual inspection [329], or by statistical analysis [330]. Here we confirmed these network motif modules in worm and plant and expanded them with the CIRc and FBUc modules (Figure 5.4). In addition, we also extended the interaction set by integrating miRNA-mRNA, regulatory and homologous interactions. Next to this, we showed that the aggregation of different composite network motifs (ALLc) can provide useful functional insights (Figure 5.5). The fact that these network motif modules are detected in two unrelated species, and comparable patterns have been detected in yeast, suggests that these topological patterns might be universal throughout GRNs in all species. The network motif modules can be linked to specific functions in GRNs and by integrating gene expression data, we revealed the dynamics of these network motif modules during development or upon stress. Through the superview analysis, in which we connected the different network motif modules with one another and with regulators, we discovered novel functional and regulatory relations between modules in the integrated GRNs context. This really demonstrated the power of our data-integration framework, since genes and regulators were found to be interacting in novel, previously unstudied, biological contexts. Higher-order organization like these network motif modules has also been observed in non-molecular and non-biological networks [331]. Here, we have provided a framework to study integrated GRNs in higher eukaryotes through network motif modules.

5.5.2 Evolution of integrated GRNs

In this study we used phylogenetic decomposition to study the evolution of integrated GRNs and the incorporation of novel genes [333]. The resolution of the age group split is dependent on the availability of genome information of the different taxa. For *A. thaliana* we are able to get a refined classification supported by multiple species in most age groups starting off from the Green Plants for almost all protein-coding genes in the integrated GRNs (Suppl. Fig. 12) [1]. However, some gaps in the taxonomy still need to be filled e.g.

ferns. For *C. elegans*, the availability of genomes in 'older' taxonomic groups is much sparser, which results in larger gaps between the different age groups (Suppl. Fig. 12). However, for *C. elegans* the phylogenetic composition goes all the way down to Cellular Organisms I.e. Bacteria. Furthermore, only 61% of all protein-coding genes in the integrated GRNs of *C. elegans* could be classified in evolutionary age groups. Hence, the study of the evolution of the worm integrated GRNs is on only part of the networks. We found the interactions to be mainly distributed over the age groups containing the most genes, which included the oldest age groups in both species.

Several methods have been used to investigate interaction age preference in biological networks. One of the first studies characterized the age-dependent evolution of yeast protein-protein interaction networks based on the interaction density of the networks, which measures the numbers of observed over expected edges between nodes of paired age groups, normalized for the size of the network [340] (Supplementary Data). The interaction density is an intrinsic property of biological networks, but in order to infer preference patterns a comparison to randomized networks with conserved degree distribution and conserved age distribution, is needed [334]. Other studies compared the observed number of interactions in the actual networks to the expected number to occur by chance in random networks that preserve the degree distribution of each age group [336, 338]. An intuitive view on interaction age preference is obtained by counting the edges between nodes of paired age groups and comparing these numbers to the ones obtained by permutation analysis of the gene-evolutionary age group assignments [337]. In order to accurately investigate interaction age preference, we applied several of the above described computational approaches and we largely obtained similar results using different measures (count analysis in Results and interaction density analysis in Supplementary Data), as has been observed previously for undirected interactions. In this respect, the preferential interaction between proteins of similar age was demonstrated in protein-protein interaction networks in yeast [338] and human [334] and for coexpression networks in *A. thaliana* [337]. However, these studies mostly used a limited number of age groups and only one type of interaction network. Using detailed phylogenetic decomposition, we showed that for undirected protein-protein interactions in *C. elegans* and *A. thaliana*, while the majority of interactions is between older and younger genes (Figure 5.7), genes preferentially interact with genes of a similar age (Figure 5.8). Similar results are obtained for all undirected interactions in *C. elegans*, hence including genetic interactions. Interactions between paralogs can only partially account for the age-dependency in the undirected networks. Overall, we can conclude that functional interactions tend to occur between proteins of similar evolutionary age. This indicates that introduction of a novel biological function involved the integration of a set of interacting genes in the GRNs. We expanded the interaction age preference to directed interactions (protein-DNA and regulatory) in both species. However, we have to take the distribution of TFs over the different age groups into account upon interpreting the results. In *C. elegans*, the TFs distribution over the age groups is shifted towards the Eumetazoa, which has more than half of the studied TFs (Suppl. Table 14). Younger TFs in *A. thaliana* are scarce and lack interaction data; in the Rosids and *A. thaliana* age group even no TFs were studied (Suppl. Table 15). With these limitations in mind, we found that regulatory TFs favoured older or same age target genes. Contrary to undirected interactions, directed interactions seem to cross the age groups as is also observed on the motif and module level. We also found that interactions with experimental binding data (physical protein-protein and protein-DNA interactions) are generally age homogeneous, while interaction types that can also be indirect (genetic and regulatory) do not show any preferential age homogeneity. Our

findings correspond to the observation that in the course of evolution of a GRN regulatory interactions are acquired much faster than protein-protein and genetic interactions [388].

Different mechanistic models have been introduced to explain the evolution of biological networks, especially protein-protein interaction networks. In the “preferential attachment” model, new proteins preferentially attach to highly connected nodes [389]. The “duplication and divergence” model states that new proteins originated through duplication, initially connect to all the neighbours of the node that has been duplicated and that connections diverge over time [390]. However, these models are not able to mimic the high modularity and the homogeneous age preference of protein interactions. In the “crystal growth” model, the network grows by anchoring and extension, where a node increases its degree either by becoming a new module (anchoring) or by extending an existing module [340]. This model incorporates the tendency of protein-protein interactions to interact within the same age group, the central aggregation of older subunits and the peripheral scattering of younger subunits and hence corresponds with our findings at the interaction level. The most recent model for protein-protein interaction evolution that mimics real protein-protein interaction networks the best, is the “network motif” model, which is based upon the fact that network motifs or protein clusters are incorporated into the network instead of single proteins [341]. It was confirmed in a yeast protein-protein interaction network that proteins of the same age class tend to form motifs, are densely interconnected, co-evolve, share the same biological function and tend to be within protein complexes [339]. Similar to the network motif model, motifs were also used as building blocks to model transcriptional networks in bacteria [391]. In accordance with these models, we looked into age patterns in the network motifs and network motif modules to get insight in the evolutionary mechanisms for GRN formation (Figure 5.9). Our age preference analysis at the motif and module levels indicated a strong age homogeneity preference for COM motifs and COMc modules and a strong age heterogeneity preference for FFL motifs and FFLc modules, especially in *A. thaliana*, which is in agreement with our results on interaction age preference of undirected and directed interactions, respectively. In *C. elegans* this is only partially true, since here we found COM motifs with at least one age homogeneous interaction more in COMc modules, while age homogeneous COM motifs are less clustered, and we did find an overrepresentation of age homogeneous FFL motifs as well. This can be explained by the dominance of the Eumetazoa and Eukaryota age groups in the *C. elegans* interactions, motifs and modules. Overall, we found the overrepresented age types in de motifs to be more incorporated in the modules. Compared to the other module types, the COMc modules were more inclined to comprise a single age group in both species (Suppl. Fig. 14). However, modules were mostly composed out of genes from the evolutionary age groups containing the most genes i.e. Eukaryota, Eumetazoa and Caenorhabditis in worm, and Green and Land plants in plant. Hence, they mostly consisted of older genes and only had a smaller fraction of younger genes. This hints to the fact that the younger genes likely attach to the older core of the network during GRN evolution. Taking into account our results at interaction, motif and module level, we postulate that novel genes attach together to the GRNs in a specific biological functional context, regulated by one or more TFs that also target older genes in the GRNs. Hence, for the undirected interactions, this is in accordance with the “network motif” model [341], although single genes might accompany the addition of network motifs and modules in GRN formation over evolutionary time, as low-connected genes are missed through data-integration based on network motifs or network motif modules.

5.5.3 Influence of gene duplication on network evolution

In *C. elegans*, small scale duplications (SSD) make up the biggest portion of the duplicates. These are frequently partial or lack the original regulatory sequences [342]. In *A. thaliana*, WGDs are the source of many duplicates, next to SSDs [16]. A particular difference between integrated GRNs in *A. thaliana* and *C. elegans* is that homologous plant TFs targeting the same genes tend to physically interact more both through protein-protein and protein-DNA interactions, but homologous interactions between TFs occur more than 7 times more in plant than worm. The faster divergence of genes after SSD in terms of divergences of sequence [227], expression [226], protein interaction partners [63] and regulatory connections [343] makes that homologous relations between TFs in *C. elegans* are no longer detected.

The differences in divergence between SSD and WGD also have an influence on the age groups classification of genes since they are categorised on the oldest occurring species in the gene family or with a shared ortholog. WGD-duplicates tend to stay within the same gene families, further expanding them, while the faster divergence of SSD-duplicates allows them to create novel gene families. This potentially explains why there is a much higher number of older genes in *A. thaliana* and why there are also more genes in the younger groups of *C. elegans* than in *A. thaliana*: 14% of the *A. thaliana* genes originated after the Brassicales split off (estimated 68 MYA ago) compared to 29%, after *C. elegans* diverged from other Caenorhabditis worms (estimated 60 MYA ago). This is reflected in the numbers of TFs in both species' age groups: 14% of worm TFs belong to Caenorhabditis or younger age groups, while only 2% of plant TFs are associated to Brassicales or younger age groups (Suppl. Table 14 & Suppl. Table 15). TFs expand through duplication, often WGD, and are retained for long periods after duplication [29, 229, 392]. In *C. elegans*, the TF age distribution is diverse, which links with the fast evolution in sequence and function, and often loss probably because of dosage balance reasons, after SSD [29, 393]. Still despite these differences it leads to an interaction pattern with no single preferential age group but its own in the undirected networks of both species.

Motifs can originate from duplication of genes. In both species we see an overlap between HPP/PPP motifs, which hints to the contribution of duplication on complex motif formation. Since this overlap is only for a very small fraction of the total amount of PPP motifs, we exclude that duplication is a major creator of motifs but still on the cluster level this overlap gives rise to star like modules around a homolog pair (e.g. Cele COMc 70, Figure 5.4). The influence of duplication on network motif formation is visible within the motifs with directed interactions in *A. thaliana*. Similar results were obtained for genes after WGD in yeast[394]. For motifs containing protein-DNA interactions, we found that homologs contribute more to motif formation in *A. thaliana* than expected. Likewise, we found a large overlap between HDD/DDD and HDD/PDD motifs in Arabidopsis and not in *C. elegans*. Homologous interactions between TFs and the overlap between motifs can explain the overrepresentation of certain age patterns in motifs. In *A. thaliana* the COP motifs are preferential of the YYO/OYM-type, where two younger TFs of the same age or an old and young TF interact and target a gene of a different age group. In the context of duplication, this could be seen as a homodimer (YYO) which becomes a heterodimer after divergence (OYM). This explains the overlap between both motif types HDD/PDD. The FFL motif DDD turned out to be preferential age heterogeneous, OYM in both species, followed by YOO and MOY in plant, and by SSS, MOY and YYO in worm. This shows that novel genes are incorporated at every position in the FFL, but also that additional regulatory layers could be generated by the doubling of one of the TFs and the gain of a regulatory interaction. The gain of regulatory layers shows that

evolution increases the complexity of GRNs, which allows adaptation and more specific regulation of downstream processes [388]. This is in correspondence with the fact that novel TFs show a higher target binding specificity in *A. thaliana* as compared to TFs of ancient families [395]. In *C. elegans*, we detected overlap between HMM/GMM, which shows that miRNAs of the same family often are genetically linked and overlap between DDH/DDP/DDG motifs, which represents interacting duplicate targets through either genetic or protein-protein interactions.

In summary, we report the presence and biological relevance of network motif and network motif modules in the integrated GRNs of *C. elegans* and *A. thaliana*. These topological patterns are potentially universal in networks of gene regulation. Depending on the interaction type being functional or regulatory, we find different interaction age preferences in GRN evolution, which are similar in both species.

5.6 MATERIALS AND METHODS

5.6.1 Source of interaction data

An overview of the molecular interaction data of the integrated GRNs can be found in Table 1. For *C. elegans*, the undirected molecular interaction data were compiled from the following resources: 9739 protein-protein interactions (P) from Wormbase WS234 [396], Worm Interactome version 8 [397], BioGRID 3.2.97 [398], and literature [347, 399-402]; 3830 genetic interactions (G) from Wormbase WS234 [396], BioGRID 3.2.97 [398] and selected publications [403-406]; 6502 homologous interactions (H), which consisted on the one hand of 6348 paralogous protein-coding genes determined by an all-against-all BLASTP of the *C. elegans* proteome WS220 (E-value < 1e-25, percent alignment > 60%) and on the other hand of 154 paralogous miRNAs with identical seed sequence identified through BLASTN. For the regulators in the directed molecular interactions, TFs were defined as in WormBook [407], while miRNAs were retrieved from miRBase [408]. The 13,747 protein-DNA binding interactions (D) consist of two types of experimental data, Y1H and ChIP. The Y1H dataset contains both large and small-scale data sets [309, 347, 349, 409-415]. The ChIP dataset was taken from modENCODE, where TF-protein-coding target gene interactions predicted from ChIP-seq data by the TIP algorithm were used with a quality score of 1 [416]. The 3948 regulatory interactions (R) comprise genes with a two-fold log2 change in gene expression upon knock-out or knock-down of the regulator [401, 417-424], supplemented with regulation associated interactions from the text-mining database EVEX [61]. The 6177 miRNA-mRNA interactions (M) entail experimental confirmed interactions from miRTarBase (49%) [425] and PicTar predictions conserved in five species (51%) [426, 427]. Gene identifiers of all protein-coding genes and miRNAs were converted to Wormbase WS245 using WormBase Converter [428] and a Perl script, only keeping the genes (and their interactions) with unchanged WS identifier or that merged/split to a new WS identifier. Finally, the worm integrated GRNs contained 43943 interactions between 845 TFs (92% of all TFs), 172 miRNAs (67% of all miRNAs) and 12,095 protein-coding genes (67% of all protein-coding genes [429]).

For *A. thaliana*, the undirected molecular interaction data were collected from the following resources: 52,613 protein-protein interactions (P) from CORNET 3.0 (experimentally validated interactions only) [263], BioGRID 3.2.97 [398], MIND (high confidence) [262] and the Arabidopsis Interactome [63]; and 5254 homologous interactions (H), which consisted on the one hand of 5226 paralogous protein-coding genes established from phylogenetic tree-based gene families [229] and on the other hand of 28 paralogous miRNAs with identical seed sequence identified through BLASTN. For the regulators in the directed molecular interactions, TFs were named by PlantTFDB 3.0 [430], while miRNAs were found in miRBase [408]. The 29690 protein-DNA and transcription regulatory interactions (D) include ChIP data from a meta-analysis of publicly available data [431] and a high confidence reference set that combines ChIP binding and expression upon TF perturbation [353], Y1H data from literature [378] [432, 433], protein-DNA binding and/or transcription regulatory interactions from AtRegNet [434] and differential expression analysis upon TF perturbation [353]. For Arabidopsis, both protein-DNA binding and transcription regulatory interactions are combined in D, since AtRegNet does not specify the type of molecular interaction or experimental method and several interactions from AtRegNet and literature involve both DNA binding and differential expression upon TF perturbation. The 2122 miRNA-RNA interactions (M) contain experimental confirmed interactions from miRTarBase (5%) [425] and psRNATarget predictions using standard parameters on the TAIR 10 transcripts (95%) [435]. In all

Arabidopsis interactions, only protein coding genes and miRNAs with a TAIR 10 gene identifier were kept. Interactions involving mitochondrial and chloroplast genes were removed. As symbolic gene names, we used the primary symbol name from TAIR. Finally, the plant integrated GRNs encompassed 89679 interactions between 1519 TFs (88% of all TFs), 174 miRNAs (41% of all miRNAs) and 19001 protein-coding genes (69% of all protein-coding genes) [436]. Self-interactions were removed in all the networks.

Table 5.1: ***Overview of the different types of molecular interactions in the integrated GRNs of C. elegans and A. thaliana*** *respectively. P = protein-protein, G = genetic, H = homologous, D = protein-DNA*, R = transcription regulatory, M = miRNA-mRNA interactions. *In the case of A. thaliana, protein-DNA and transcription regulatory interactions are combined in D, since the AtRegNet source does not specify the type of molecular interaction or experimental method i.e. protein-DNA binding or transcription regulatory interaction and several interactions from AtRegNet and literature involve both DNA binding and differential expression upon TF perturbation. Regulators indicate TFs or miRNAs.*

	C. elegans				*A. thaliana*			
Number of	edges	nodes	regulators	targets	edges	nodes	regulators	targets
P	9739	4287	/	/	52613	10266	/	/
G	3830	1823	/	/	/	/	/	/
H	6502	4807	/	/	5254	8896	/	/
D	13747	3989	603	3733	29690	12721	399	12632
R	3948	3283	70	3235				
M	6177	1499	144	1355	2122	1623	171	1452
Total	43943	13112	611	6486	89679	20694	570	13373

5.6.2 Topology of the networks

The topology of the networks was analysed in R using the igraph package [437].

5.6.3 Network motif detection and enrichment

Three-node motifs were detected by ISMA (Index-based Subgraph Matching Algorithm) [346]. Two-node motifs were detected by a Perl script (https://gitlab.psb.ugent.be/jofoo/NetworkMotifModules.git). To calculate motif enrichment, 1000 random networks with the same degree distributions as the real networks for each interaction type were constructed through an edge swapping algorithm in the Matlab Motif Clustering Toolbox [330]. The enrichment of each detected motif compared to random networks was calculated using the Z-score $Z= (N-\mu)/\sigma$, in which N is the number of motifs in the real networks, μ the average and σ the standard deviation of the number of motifs in the random networks.

5.6.4 Network motif clustering

Network motif clustering was performed by the hypergraph-based spectral clustering algorithm SCHype with standard settings [350]. The different 3-node motifs were clustered into seven different types. Next to these groups all motifs were clustered together and separately. We filtered out modules smaller than 5 nodes and bigger than 100 nodes, and modules consisting only of homologs, because they are less informative or not interpretable.

5.6.5 Functional analysis on the integrated networks

All modules where visualized together with functional data in Cytoscape. For each module, GO Biological Process enrichment values ($p < 0.05$) were calculated by the BINGO 2.44 Cytoscape plugin using the Benjamini and Hochberg multiple testing correction [29]. We used the core GO ontology release 2015-01-09 together with gene annotations files for *A. thaliana* GOC: 08/01/2016 and *C. elegans* GOC: 07/17/2016.

5.6.6 Integration of expression profile data

For the *C. elegans* microarray data, we derived expression ratios for embryonic development by dividing the expression matrix by the overall average [351] and for embryonic and postembryonic development by dividing tissue-specific expression by its whole animal reference set at a specific developmental stage [352] (Suppl. Table 7 and 8). The Arabidopsis abiotic stress-dedicated microarray expression profiles consisted of expression ratios in 199 experiment over control conditions [353] (Suppl. Table 9). Coexpression within modules was calculated by the average Pearson Correlation Coefficient of all the genes in a module (nPCC) and the p-value from the Z-score upon comparison to 1000 random modules of the same size picked from all clustered genes (Atha: 20695 genes, Cele: 13112 genes). For *A. thaliana*, dynamicity of the modules was analysed by the Expression Correlation Differential Score (ECD), which sums up the differences between the Pearson Correlation Coefficient (PCC) in abiotic stress and control conditions for all the edges in the module, a measure that was originally developed for motifs [438]. Therefore, stress conditions were grouped per abiotic stress type (Suppl. Table 9). Since the calculation of the PCC requires multiple replicates for a specific experimental condition, we were only able to calculate the ECD for plant, and not for worm. The PCC of every module gene was calculated in the environmental stress, as well as in the control conditions. The ECD was then calculated by the following formula: $ECD = \sum_{all\ module\ edges} abs(PCC_{stress/development} - PCC_{control})$.Finally, the significance of the ECD was analysed by comparing the ECD in the real module versus the ECD of 1000 modules with the same number of genes through permutation.

5.6.7 Superview

The super view representations of the networks were created using all modules of size 5 to 50 nodes. Modules sharing 50% or more of their genes were merged under the name of the biggest module. We counted the number of interactions going from a gene in one module to a gene in another module for each interaction type separately. This results in the total number of interactions between two modules. This observed number of interactions between modules was compared to the number of interactions between 1000 random modules with the same sizes as the original modules. The random modules were obtained by randomly selecting genes from all genes present in the modules of the integrated GRNs. A Z-score and p-value were calculated to compare the observed versus the expected value. To assign regulators to modules we integrated sets of regulatory interactions (D/R/M) with the modules. For this we counted the number of genes in a module regulated by a certain regulator. Regulators that were already in the module were not counted. We compared this count with the number of regulatory interactions going to the random modules with the same size. A Z-score and p-value were calculated to compare the observed versus the expected value.

5.6.8 Visualization

All network figures where made using Cytoscape. For the interactive web visualization, a custom version of CyNetShare was used (http://idekerlab.github.io/cy-net-share/). A standalone Java tool called ModuleViewer visualized the expression ratios together with other relevant biological data into customized heatmaps [353].

5.6.9 Phylogenetic decomposition

We applied phylostratigraphy to derive the evolutionary origin of the genes [333]. Specifically, *A. thaliana* gene families were assigned phylogenetic ages based on the oldest lineage that still contains an ortholog of the gene family i.e. the earliest common ancestor of the gene family. As an example, if a gene family contains 4 genes from species in the Brassicaceae lineage and one gene from Physcomitrella patens, it is classified as Land plants/Embryophyta. Orthologous gene families were downloaded from PLAZA 4.0 dicots (http://bioinformatics.psb.ugent.be/plaza/versions/plaza_v4_dicots/) [1]. They are constructed out of 55 fully sequenced species with a wide distribution over the different lineages. This resulted in 10 age groups: Green plants, Land plants, Vascular plants, Seed plants, Flowering plants, eudicots, Rosids, Brassicales, Brassicaceae, *A. thaliana* (Suppl. Fig. 12). For *C. elegans* genes, phylogenetic ages were assigned according to consensus gene-age labels that are based on 13 orthology inference algorithms [439], as well as Caenorhabditis genus-specific and Caenorhabditis elegans species-specific gene labels [440]. Where both methods differed, we used the oldest classification. This resulted in seven age groups: Cellular organisms, Eukaryota, Opisthokonta, Eumetazoa, Ecdysozoa, Caenorhabditis and *C. elegans* (Suppl. Fig. 12). For both species, all genes with their phylogenetic classification are listed in Suppl. Table 19.

5.6.10 Interaction homogeneity and age preference

For the age homogeneity analysis, we compared the observed number of interactions between the genes in same age groups to the expected number of interactions based on 1000 randomized networks with the same age and degree distribution. Based on this comparison, a Z-score and p-value was calculated with multiple hypothesis testing correction (Benjamini-Hochberg). For the count analysis, the observed number of interactions between the genes in the age groups was compared to the expected number of interactions based on 1000 randomized networks with the same age and degree distribution. Based on this comparison, a Z-score and p-value was calculated with multiple hypothesis testing correction (Benjamini-Hochberg).

5.6.11 Age pattern analysis in network motifs and modules

Each motif was assigned to one of 13 motif age types (Figure 5.9A). Redundancy through internal symmetry within these types was removed by selecting only the motif where the nodes are in decreasing age order e.g. for COP motifs, the motif age type OYO is the same as YOO, but OYO is chosen over YOO because there the nodes are ordered from old to young. We calculated the motif age preference by computing a Z-score and associated p-value with multiple hypothesis testing correction (Benjamini-Hochberg) of observed motif age patterns compared to expected by 1000 permutations, where nodes are shuffled for each of the three node positions separately, so the age distribution per node position is preserved. For the module age preference in each module type, we subtracted the percentage of motif age type motifs belonging to a certain motif type from the percentage of motif age type that was clustered in the corresponding module type. Formula: (relative fraction of clustered motifs) - (relative fraction of motifs in total of a certain motif age type per

module type). For example, within the COM motifs of *A. thaliana* 21% is age homogeneous (SSS) while in the clustered motifs in the module set COMc 42% is age homogeneous. This results in a relative difference of +21%, which means that age homogeneous COM motifs are preferentially clustered. To test significance, a hypergeometric test was performed with multiple hypothesis testing correction (Benjamini-Hochberg).

5.6.12 Computational resources

All data processing and data analysis was done using a combination of Perl, Python and R scripts. All scripts can be found on (https://gitlab.psb.ugent.be/jofoo/NetworkMotifModules.git). The complete search for all possible motifs in the real networks with ISMA was run on a single Linux computing node (2.4GHz, Intel) and took 3 minutes and 52 seconds and used at maximum 250Mb of memory (Script: runISMA.pl). The ISMA running time for individual motifs was between 19 and 506ms, depending on the size of the network. The same script was run in parallel for the detection of motifs in the 1000 random networks and gave a similar performance. The clustering of motifs with SCHype was run on one computing node. The running time and memory usage is shown in Table 5.2.

Table 5.2: ***Running time and memory usage of the SCHype clustering for Arabidopsis clusters****.*

Atha clusters	Running time (mm:ss)	Memory peak
FFLc	02:28	596.625Mb
COMc	05:28	2.174Gb
COPc	00:23	464.066Mb
CORc	01:12	599.223Mb
FB2Uc	00:05	149.879Mb
ALLc	21:29	3.744Gb

5.7 SUPPLEMENTARY INFORMATION

5.7.1 Availability online

A dynamic visualization of all modules can be found on (http://bioinformatics.psb.ugent.be/supplementary_data/jofoo/networks/). The source code of the computational data-integration framework can be found on https://gitlab.psb.ugent.be/jofoo/NetworkMotifModules.git.

5.7.2 Supplementary data, figures and tables

https://floppy.psb.ugent.be/public.php?service=files&t=f5be16e99f3c3f7a0e9dddc6868db29b

6 DISCUSSION & FUTURE PROSPECTS

"In order to attain the impossible, one must attempt the absurd."

Miguel de Cervantes

6 DISCUSSION & FUTURE PROSPECTS

6.1 SURFING THE DATA WAVE IN THE OMICS ERA

The data tsunami of the -Omics era is hitting the shores and makes an overload of data available which calls for new data integration, data analysis and data interpretation methods. We need to rethink how we make use of this overwhelming amount of data and explore novel research opportunities.

In the first research chapter, we made use of 37 angiosperm plants for which a good assembly and annotation was available to study gene loss and retention patterns (chapter 2). This group had a good taxonomic clade distribution which allowed us to discover general patterns in the evolution of gene families after duplication. In the meantime, the number of available plant genomes has more than quadrupled and assemblies and annotations have improved for numerous amount of species [1]. After detecting these consistent retention patterns across angiosperms, we explored genome, transcriptome, proteome, and interactome data form *A. thaliana*, *S. lycopersicum* and *Z. mays* to find out why these patterns are found. We mainly studied the relationship between the conservation of PPI and the evolutionary and functional fates of gene duplicates within different plants (chapter 3).

In Chapter 4 we presented, a novel fast and easy transcriptomics approach using 3' UTR sequencing, TranSeq. The method allows to detect the expression of the whole transcriptome using a much smaller library size, which allows to load more samples on the same line. This makes it cheaper to explore a wide range of conditions. The depth of sequencing makes it also possible to use these reads for annotation improvement.

Apart from gene loss and retention we also studied the structure and evolution of interaction networks in the model organisms *A. thaliana* and *C. elegans* (Chapter 5). There we made use of the emerging interactomics data quantity available through large scale experiments, consortia and databases (e.g. [62, 63, 256, 426, 434, 441]). For the evolutionary analysis we used a phylogenetic decomposition method which splits the genes up into age groups based on the occurrence of orthologous genes within species of older taxonomic clades. As such, this method depends on the availability of whole genome sequences. For plants a broad taxonomic range of genomes is available with gene families covering most clades thanks to the PLAZA database [1]. Still some taxonomic clades aren't covered (e.g. fern) and the more extant lineages are thinly sampled (e.g. gymnosperms). For C. elegans, the availability of genomes in distant taxonomic groups is much sparser, which results in larger gaps between the different age groups. While analysing the *C. elegans* and *A. thaliana* datasets more interaction data, novel methodologies and new data types became available. Our method using network motifs and network motif modules can be used as an integration method to study structure and evolution of biological networks as well as for other networks.

Overall, the emergence of large publicly available biological datasets has allowed us to study evolution of gene duplicates and networks. These are just the first waves, coming before the big tsunami. We need to be careful that data quality is being maintained within public databases. A lot of 'draft' genomes are being published which are of poor quality with lacking annotation, meanwhile, interaction databases contain increasingly large numbers of false positives. With the transfer of data across species, errors are also transferred. This could lead to erroneous analyses and false positive discoveries.

6.2 DUPLICATE LOSS AND RETENTION

6.2.1 Gene loss and retention patterns in core gene families.

In chapter 2, we described the similar gene loss and retention patterns in core gene families across angiosperms based on 37 angiosperm genomes which harbour a large number of shared and independent WGDs events with varying age. Three groups of gene families were identified based on their tendency to maintain duplicates: single, intermediate, and multi-copy. We observed that most core gene families revert quickly to single-copy status following duplication, which could be linked to negative effects of gene duplication such as strict expression constraints associated with increasing the absolute gene dosage [152], pleiotropic negative effects on fitness due to accumulation of mutations in duplicates [33, 43, 153, 154] or cytotoxic effects (e.g. protein misfolding) [156]. The intermediate group is composed out of putative dosage-balance sensitive genes [29, 134]. Dosage balance is thought to constrain gene divergence and contributes to the prolonged retention of genes [48, 52, 442]. Over time the constraints soften, leading finally to the loss of the duplicates. In the multi-copy group genes are maintained 'indefinitely': initially dosage balance might play a role, but the indefinite retention is through other mechanisms, for example sub-/neofunctionalization which is thought to be a slow process [157].

Non-core gene families

As we focused only on core gene families in the search for duplicate retention patterns, it is possible that we missed differences between species. Detailed cross species comparisons could reveal lineage or species-specific expansions after SSD and WGD. For example, comparisons of a species subset showed different duplication retention patterns between species following WGD [48, 125, 126, 228]. The pattern of loss and retention we discovered is biased towards WGD duplicates. In parallel, non-core gene families which are preferentially duplicated through SSD are also preserved cross-taxon [146, 147, 227, 238]. A more recent study showed that both duplication modes contribute to the biased gene retention patterns in plants [247]. This shows retention is dependent on the duplication mode.

6.2.2 Differences between duplication mode might be linked to their evolutionary contribution

Functional and divergence differences between duplication modes

In *A. thaliana*, WGD and tandem duplicates make up the largest fraction of duplicates [5]. Next to these, retroduplication and duplication through transposable element make up for a significant number of duplicates. For 30% of the duplicates the duplication mechanism is still unknown, due to the lack of signatures like synteny, proximity, and repeats [5]. In the context of divergence of duplicates, we looked at tandem (SSD) and block (WGD) duplicates. We confirmed the observations that tandem duplicates diverge faster than block duplicates in terms of sequence, expression and interaction partners (Chapter 3) [228]. Other SSDs than tandems, diverge faster in terms of sequence and expression compared to tandem duplicates within *A. thaliana* and rice [227]. The differences between SSD and WGD have next to plants been observed in other

species including human [443] and yeast [232]. From this it seems that these observations can be seen as general features of the duplication modes.

The difference in conservation of duplication modes is linked to their contribution to genomic novelty and adaptation. The faster evolution and loss of the tandem duplicates limits their evolutionary potential. Tandem duplicates need to be quickly of use, otherwise they just get lost. This might explain why tandem duplications are often found in categories related to stress where they can have an immediate effect [243]. The slower divergence of WGD duplicates might explain their higher abundance. Block duplicates are created with conserved stochiometric balance between the duplicates [110, 244, 442]. Their longer retention creates potential for evolutionary innovation and adaption [162]. This is supported by the divergence of expression between duplicates in *A. thaliana* by [444]. They reported that gene categories enriched for SSD (extracellular transport, signal transduction, stress response and transcription) have the highest expression divergence and the WGD enriched categories (cellular and developmental processes such as energy pathway, protein metabolism, intracellular transport, DNA and RNA metabolism, and cell organization and biogenesis) have the lowest divergence [444]. Putting these observations together potentially shows that SSD duplicates make adaptions in a short timeframe where WGD duplicates make adaptations on the long run possible. The extra genetic material provides a buffer which possibly helps to deal with changes in environmental conditions. It has to be noted that the contribution of tandem duplicates might be underestimated due to annotation errors [5]. Tandem duplications can be seen as a repeat in the genome assembly and therefore be annotated as a single gene (e.g. SEC10 in *A. thaliana* [246]). This can especially be a problem for draft quality genomes.

WGD are of major evolutionary importance in plants

Flowering plants are the ideal model organisms to study the effect of WGD-events. They underwent at least two WGD events and a lot of them have additional lineage or species specific WGD events. Many plant species also comprise mixed populations of diploid and polyploid individuals [107].

The longer retention of WGD duplicates might create time for evolutionary innovation and adaption and might explain the evolutionary importance allocated to WGD-events [16, 24, 445]. WGD/polyploidy events are associated with mayor evolutionary adaptations. For example, allopolyploidy events in grasses seem to have led to the dominance of C4 grasses over C3 grasses and the worldwide expansion of them [245]. Polypoid species are reported to have a higher tolerance to stress conditions (e.g. salt: [163], cold: [446], drought: [447]). WGD events have also been associated with mass-extinction events [16, 21, 22, 24, 448, 449] (figure 2.5).

Observations from plants might be used to study a wide range of human diseases

Within this thesis we studied duplication within plants, but observations made here can also be used in other species, including human, in which two WGD events have been detected that have taken place during early vertebrate evolution [18]. The link between duplicability/function and the difference between SSD/WGD (see 6.2.1 and 6.2.2) are also observed in human. WGD duplicates in humans are enriched for essential and pathogenesis related functions when compared to the whole genome, also showing signatures of dosage balance and resistance to copy number variations [237, 443, 450]. Within the same set of pathogenic genes SSD duplicates are underrepresented [451]. Variations in duplications and copy number have been linked to a wide range of disorders (e.g. cancer, ADHD, …) [451-454]. Polyploid animal cells and fungi are also reported

to have a higher tolerance to stress conditions in [162, 241]. It seems that despite the numerous differences between humans and plants, the research on angiosperm gene loss and retention of duplicates and related dosage issues might help us to understand human diseases.

6.2.3 Duplicability is linked with function within and outside the plant kingdom.

Gene duplicability is highly associated with gene function, as shown by the different GO enrichment in three retention groups (figure 3.6). The single-copy genes are biased towards essential genes, functioning in genome integrity pathways and organelles while multi-copy genes are biased towards functions involved in interactions with the environment in particular transporters, signalling transducers and cell communication [29, 49]. The duplicability patterns do not appear to be limited to the plant kingdom: in a study focusing on the duplication history of genes across 17 fungi genomes, a similar functional separation was observed [230]. Likewise, a large-scale analysis of prokaryotic genomes suggested that the number of genes functioning in DNA repair and replication remains relatively constant irrespective of genome size, whereas the number of TFs, genes involved in signalling and transporter genes, seems to increase with bigger genome size [149, 150]. Consequently, patterns of duplicate retention and loss for core genes in angiosperms and other organisms appear to abide by these general function-based rules.

Despite the strong correlation between gene duplicability and gene function, it remains to be further investigated which evolutionary mechanisms are responsible for the observed strong bias in duplicate retention patterns. It remains to be established whether gene function directly influences gene duplicability or whether biased gene retention could be a by-product of other evolutionary phenomena instead, for example the preservation of intermolecular interactions (dosage balance) or sequence constraints related to high levels of gene expression [96, 170]. In chapter 3, we made the first steps to investigate the influence of interactions in a cross-species comparison. The data from this chapter could be also be used to study the link between interactions involved in certain processes and duplicate retention/divergence. The first results show that the retention is biased towards the interaction properties rather than towards the function.

6.2.4 Influence of protein-protein interactions on duplicate retention and the role of dosage balance sensitivity

The dosage balance theory has been put forward to explain the differential retention of genes after duplication. This theory explains why duplicates are thought to be maintained after WGD, but not after SSD [38, 442]. The connections in the gene interaction network are essential to this. We observed that longer retained duplicates (intermediate and multi-copy) are enriched for PPI (Chapter 2 & 3), and investigation of the influence of protein-protein interactions on duplicate retention revealed that PPI restrain sequence and expression divergence in block and tandem duplicates with similar patterns between *A. thaliana*, *S. lycopersicum* and *Z. mays* (Chapter 3). Comparison with the modelling results of Tasdighian et al. showed that genes with PPI are overrepresented in genes which are thought to be dosage sensitive based on their reciprocal retention pattern [238].

We investigated multiple species by transferring the PPI from Arabidopsis through gene families to other species. This allowed us to show similar trends for *S. lycopersicum* and *Z. mays* as for *A. thaliana*. Due to the

transfer the resolution of the PPI network is lowered. False negatives occur because of the lack of experiments or because of divergence (genes are located within separate gene families) and false positive occur due to the transfer to all genes in the same gene family. This results in an overestimation of duplicates which both have PPI. Similarly, carefulness needs to be applied when using interactions form databases, large scale experiments often have a high false positive rate [455] and interactions are transferred through orthology with removal of the differences between duplicates [456]. To overcome this further study is needed to characterize the interactions, preferentially with multiple experimental techniques. For regulatory interactions a similar approach as for PPI was attempted in order to see if there is a link between experimental protein-DNA interactions and duplicate retention (results not shown). Due to the lack of experimentally validated interactions no conclusions could be made.

All the results provide indirect evidence that dosage sensitivity due to interactions is playing a role in the retention of duplicates, but experimental validation is still needed. A first step in this might be improvement of the interactome, by expanding the network to all genes both for protein-protein interactions and protein-DNA interactions and by filtering out the false positives. In an ideal situation we would know the whole interactome for multiple angiosperm plants with a diverse timing in WGD events, which would allow us to investigate the influence of interactions in an evolutionary context. This could help in answering remaining questions like: “How is gene retention linked to the network context?”, “Is there a difference between transient and stable interactions?”, “Do specific protein domains influence dosage balance sensitivity?” and “What is the influence of protein complexes?”.

In angiosperm plants the involvement of complexes in dosage balance is not studied in detail due to the lack of well-defined plant protein complexes. Most of the complexes are transferred from other species like yeast and human. In those species several studies have focussed on the link between duplication and protein complexes, especially the topology of complexes was found to be important and could be linked to expression differences between duplicates [457-460]. In plants a similar approach as [457] could be used. They used TAP to study the composition and evolution of a wide set of yeast complexes.

To study actual dosage balance between genes, proteomic quantification could be used to show if the actual dosage of genes is linked to a phenotypic effect. This could be achieved by characterizing knock-out mutants of one copy of the duplicated gene pair for each of the different retention groups (single, intermediate or multi-copy group). Similarly, a duplicate copy could be inserted for a gene in the single copy group. Within populations a lot of natural polyploids are occurring [446, 461, 462], comparison of the gene dosage in both of them could show how the dosage in plants responds to a WGD-event. It would also be interesting to study the involvement of copy number variations. In Arabidopsis and rice for example a lot of sequenced accessions are already available [2, 3]. Apart from experimental approaches, modelling approaches could be used to answer the questions. Up till now, models still fail to realistically model transcriptional regulatory interactions, gene duplication, and duplicate divergence [463]. To make statistical thermodynamics modelling accurate a large set of thermodynamically well-defined interactions would be needed.

Dosage balance is not the only theory for longer retention: it is only part of the complex puzzle of gene and genome evolution [240]. Apart from looking at the protein level, we could look for clues of gene loss and

retention on the DNA level by studying remnants of lost genes (see 6.2.7), and on the RNA level the transcriptional silencing of genes could be studied (see 6.2.6).

6.2.5 Molecular characteristics of gene families determine fate of duplicates across species

We have shown strikingly similar loss and retention patterns across angiosperms, differences in duplication mode retention, influence of interactions and the link between function and retention (chapter 2 & 3). None of these observations are independent of each other: duplication mode shows a link with function and interaction preference, and interactions are linked with the function of genes. Most of these observations can be seen as characteristics of gene families and are also found within a wide range of other species. All of these factors contribute to the patterns found across species, but they can't explain the full mechanism behind this. They all work together in a complex system shaped by evolution. Further study is needed to find out why and how certain families are retained and others are lost. Due to the broad species ranges in which we detect the same results, we might start thinking there is a universal mechanism shaping the genomes of all species.

The results suggest that duplicates within the same gene family evolve at a similar rate and potentially point out that the loss and retention mechanisms are dependent on the molecular characteristics of the gene family. For this it would be interesting to study the loss of genes cross species. The genomic, transcriptomic and interaction data from chapter 2 and 3 could be used to study a diverse range of hypothesis. For example, it is possible to test if duplicated genes which are lost in one species have a higher sequence/expression divergence or one copy is transcriptionally silenced in species in which they still are present in duplicate. It is also possible to test, if one of the interacting duplicates is lost, the other interaction partner also gets lost. The latter is similar to observations made in yeast [48]. Finally, the data also allows to study how the duplicates from the shared WGD events (e.g. eudicot WGD) evolved and potentially resulted in different functions.

Within this thesis we only studied a limited number of factors influencing duplicate loss and retention in gene families. For example, the link between regulation and the dosage of genes is not fully understood. It is easy to think that instead of removing or altering the function of the gene itself, it is much easier to alter the regulation of a gene. A concrete example of this can be changes in the cis-regulatory elements that cause non/sub/neo-functionalisation of one copy. Detailed study of regulatory differences between duplicates could answer how they are influenced by changes in regulation [343, 464-469].

Another example is the epigenetic involvement, which has been shown to influence retention and silencing of genes [54, 470]. Epigenetic marks could help the resolve dosage issues, by silencing duplicates which are prone to paralog interference or by inducing condition specificity of duplicates. The influence of epigenetics has mostly been observed between the different subgenomes in allopolyploids [145, 471, 472]. Studying this at the gene family level might reveal the biases towards a specific set of them. Next to that it might be interesting to study how epigenetics evolves over time in autopolyploids [473].

Alternative splicing produces multiple isoforms of a gene which increases the protein diversity [474]. This is similar to gene duplication as they can be seen as inparalogs [475]. The link between duplication and alternative splicing is not fully elucidated. For example, after duplication isoforms can be divided among duplicates creating sub-functionalisation [476]. In human and mouse duplicated genes were found to undergo less frequently alternative splicing [477], while in rice the opposite was observed [478]. Potentially studying alternative splicing patterns across angiosperms could show a link between alternative splicing and duplicate loss and retention patterns in the gene families. Also copy number variants could be an interesting factor to study at the gene family level (see 6.2.2).

6.2.6 Duplicate detection on the RNA level

The loss and retention patterns show a still ongoing loss of gene duplicates. In order to test hypothesises like dosage sub-functionalisation [35] and reciprocal silencing [466], the transcribed set of genes within different angiosperms was investigated. This research was started with microarray expression data, because of the wide range of experimental conditions and species available and the easy accessibility through online tools such as CORNET [202] and MORPH [479]. Microarrays have problematic detection of duplicates, due to low detection rate, cross hybridisation between close homologs and a limited set of genes for which there are probes [480, 481]. Within this set of missing/not-detected genes often recent duplicates are found. Still microarrays have proven to be useful for duplication research [226, 227],

In chapter 3 we used RNAseq experimental data with unique read mapping, in order to abbreviate some of these issues. RNAseq is not limited to a probe set, which theoretically allows for the detection of all expressed transcripts. We collected large RNAseq compendia for *A. thaliana, S. lycopersicum* and *Z. mays*. This allowed us to uniquely detect all copies for the majority of the duplicates (Chapter 3). Still there was a big difference in detection between the duplication modes. Of the slower diverging WGD duplicates more than 80% was detected, while only 30-40% of tandem duplicates was detected. This is either due to silenced expression or due to non-unique mapping because of little sequence divergence.

In chapter 4 we presented a novel 3'end sequencing approach, TranSeq, which has cheap and fast library preparation. By sequencing only the non-coding 3' end, the method enables detection of all gene within the genome using a smaller library size (Figure 6.1). This smaller library size allows to sample more conditions or sample more in depth for a lower price. Refinement of the technique is still needed to make it completely comparable to classical RNAseq in terms of differential expression. Next to expression detection rate TranSeq is also improving genome annotation by allowing better determination of the end of genes, which should also improve the mapping of classical RNAseq reads.

Despite the better detection rate of duplicates, a large portion of duplicates with low Ks values were still not detected (uniquely) in tomato using classical RNAseq. The possibility to detect genes in depth and the broad sampling with TranSeq, combined with the higher variability of the 3' UTR, should enable to discriminate effectively between genes within a gene family. This could help to detect if duplicate copies are silenced or active under conditions specific. Detection of the duplicate genes is especially challenging in plant species with high ploidy level which are interesting in agriculture (e.g. hexaploidy wheat and octoploid strawberry). It is important to detect if there is a dominant copy within a gene family, if some copies are expressed in rare

conditions or if gene duplicates lost their ability to transcribe (e.g. pseudogenes). The discrimination between genes in a gene family can help to detect sub-functionalisation between duplicates and can potentially be linked to certain traits.

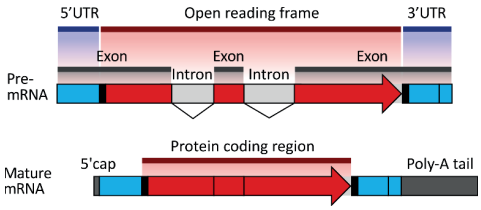


*Figure 6.1: **Structure of Pre-mRNA and Mature RNA**. Figure adapted from [482].*

6.2.7 Pseudogenes: Archaeology and future of the genome

Most duplicate genes are thought to be lost through pseudogenization. These pseudogenes lose their original function and their ability to be transcribed or translated. This doesn't mean they are without function. Pseudogenes have been studied within human and mouse genomes, where they are sometimes found to take up a regulatory role. This is often associated with the emergence of RNA structures like non-coding RNAs and are linked with diseases in human (e.g. cancer) [483-486]. Within plants genomes pseudogenes have been explored in only a few species without looking into detail whether or not they are still functional [487-490]. Current methods are inaccurate and lack sensitivity in plant pseudogene detection [488, 491-493] (reviewed in: [490]).

Studying pseudogenes systematically as remnants of gene duplicates could be useful for finding the gene loss mechanisms of duplicate genes. More specific, it is possible to look within aligned co-linear genomic regions for pseudogenes on the homeologous chromosome regions (Figure 6.2). This could be done by using the multiplicon regions found within PLAZA [251]. Since these regions originate from WGD, remnants within these regions can be linked to it and could tell more about the ongoing loss of duplicates that originated during these events. Studying these remnants might also reveal functions for them, for example they can give rise to micro RNAs or non-coding RNAs which can take up novel regulatory functions [483]. They can go through a cryptic phase where they are allowed to gather a lot of mutations which enables them to perform novel functions if they are switched on again [483]. These genes could appear as novel genes in the network. The detection of whether duplicates are still expressed is also challenging as discussed in section 6.2.6.

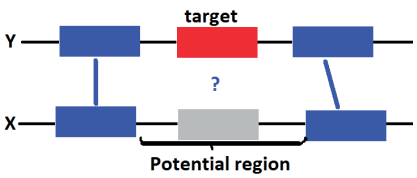


*Figure 6.2: **Detection of remnants/pseudogenes derived from WGD.** Two collinear regions (Y and X) with corresponding block duplicates (blue). Pseudogene in the X region can be detected by Blasting the gene in the other region (red) against the homeologous region where the gene has been lost (grey).*

The genomic information which is available now, enables a systematic approach for a wide range of sequenced angiosperms. It would be possible to detect the different loss mechanisms and their contribution (e.g. immediate knock out by transposon insertion, gradual decay by mutational load, promotor degradation

or possible other mechanisms). This can show how the gene families return to their preferential copy number state, the speed of duplicate loss rate in different angiosperms, how novel functions emerge, how WGD are resolved over time, how the similar retention patterns found in angiosperms emerge. It can also help to improve genome annotation, by detecting previously unknown genes of functional components. To summarize: by digging up remnants from evolutionary events, we can possibly predict what the future will bring.

6.3 GENE REGULATORY NETWORK STRUCTURE AND EVOLUTION

Chapter 5 proposed a framework to study integrated GRNs in higher eukaryotes through network motif modules. The unbiased method does not favour any interaction type or experimental methodology over the other, and preserves the identity of the interaction type as compared to other data integration methodologies that perform benchmarks using true positive data sets, GO or KEGG [380-382]. The integration of complementary data types through 2- and 3-node motifs offered advantages and provided useful insights in the study of gene regulation and in GRN evolution. Motifs, such as the well-described feed-forward loop, connect the regulatory levels (transcriptional and posttranscriptional) and integrated the directed and undirected interactions into easily interpretable patterns of gene regulation. Motifs aggregate into network motif modules that can be linked to specific functions in GRNs by integrating functional and gene expression data. Through the superview analysis, in which we connected the different network motif modules with one another and to regulators, we discovered novel functional and regulatory relations between modules in the integrated GRNs context. This demonstrated the power of our data-integration framework, since genes and regulators were found to be interacting in novel, previously unstudied, biological contexts. The network motif modules expand the 'themes' discovered previously in yeast [329, 330]. The fact that we found them in two unrelated species, *C. elegans* and *A. thaliana*, might hint that these structures are universal for biological interaction networks. Higher-order organization like these network motif modules have also been observed in non-molecular and non-biological networks illustrating its broad applicability [331].

The success of the integration method is dependent on the availability and quality of interaction data. Large-scale Y2H screens are known to have a high false positive rate [455] and only a fraction of the binding events detected by ChIP has a regulatory effect [494]. The quality of the interaction data can be improved through confirmation by multiple experimental techniques. For example, protein-DNA interactions can be improved by combining the binding experiments with analyses of the transcriptomic effect of knocking out the TF [494]. Also, by integrating interactions through network motifs and aggregating those, the influence of false positive interactions is reduced, yet this also removes the lowly connected genes. We studied the plant *A. thaliana* and the worm *C. elegans* which have sufficiently available data, although a lot of genes remain unstudied. The dependency on data availability makes it hard to study non-model organisms. While interactions are poorly conserved between species, the module level is claimed to be the most conserved level across species [387]. Using modules for translational research makes it possible to transfer the knowledge from the integration method to other species.

Apart from the data quality and availability, the cellular context and condition specificity is also important. As we started from static networks and integrated the expression data in the end of the process, it is difficult

to identify condition specific modules. Ideally conditional or temporal interaction networks should be used to generate motifs and network motif modules. These networks can be approximated using single cell transcriptomics or proteomics data [90-92].

6.4 EVOLUTION OF INTEGRATED GRNS

We made use of the structural network components and phylogenetic decomposition to study the evolution of integrated GRNs, the incorporation of novel genes [333] and the influence of duplication on this. We concluded that functional interactions tend to occur between proteins of similar evolutionary age and that interactions between paralogs can only partially account for the age-dependency in the undirected networks. Contrary to undirected interactions, directed interactions seem to cross the age groups. Taking into account our results at interaction, motif and module level, we postulate that novel genes attach together to the GRNs in a specific biological functional context, regulated by one or more TFs that also target older genes in the GRNs. Hence, for the undirected interactions, this is in accordance with the "network motif" model [341]. Although single genes might accompany the addition of network motifs and modules in GRN formation over evolutionary time, as low-connected genes are missed through data-integration based on network motifs or network motif modules.

The currently available data for *A. thaliana*, *C. elegans* and other species allows to cover a much wider range of processes and conditions and should provide a more detailed analysis. Especially the directed interactions would profit from an expansion. An update of the expression compendium to a RNAseq compendium could help to get a better view on duplicate divergence in the GRN context (see 6.2.6 Duplicate detection on the RNA level). Combining this with a more detailed sampling of the phylogenetic tree should allow the determination of contribution of novel genes at specific timepoints in evolution. Networks from multiple species with a shared history could be compared to show how the same genes are incorporated in the network of different species. We observed that genes in interactions with experimental binding evidence are preferentially age homogenous (protein-DNA and protein-protein), while interactions without experimental binding are age heterogenous (regulatory and genetic). From this we can deduce that co-expressing and computational predicted networks should not be used as a proxy for network evolution in interaction networks.

The topological components of GRNs (network motifs and network motif modules) could be combined together with the observations in undirected and directed network evolution in a model to simulate the evolution of integrated GRNs. This type of modules could also be useful to study gene loss and retention in a network context (see 6.2.4).

6.5 THE NEXT STEP AND GENERAL CONCLUSION

In the first part of this thesis we used genomics, transcriptomic and interactomics data to reveal uniform retention patterns across the flowering plants and studied different factors that may be influencing this pattern like gene function, duplication mode and protein-protein interactions. In the second part, we presented a cheaper and faster RNA sequencing procedure based on 3'UTR sequencing. In the last part we showed topological organisations and evolutionary patterns in *A. thaliana* and *C. elegans* gene regulatory networks. Combining the gathered knowledge, opens opportunities to study the loss and retention patterns in more detail.

Therefore, as a follow up on this work, I think it is interesting to study the expression divergence between duplicates in different species. Ideally this could be done by compiling a TranSeq compendium for a representative in different angiosperm subgroups (e.g. Asterids, Fabids, Malvids and Monocots). For this, TranSeq should first be further optimised to determine the ideal library size for the detection of gene duplicates and the optimal read mapping algorithm should be determined. This set of compendia would allow to perform cross species comparison of expression divergence and answer questions like. Do duplicates which are lost/diverged in one species also diverge faster in other species? Is this linked to the observed loss and retention groups in the other angiosperms? Do duplicates from a given gene family evolve at similar rates in different species? Is there a dominant copy and is that copy always the same in different species that share a WGD? Do we observe dosage sub-functionalization in certain functional groups or specific protein classes [35]? Next to these questions the TranSeq dataset could also be used for the improvement of gene models and to study the link between splice variants and duplicates in the selected species.

In a second phase it would be interesting to study the regulatory divergence of duplicates in this species set and link this to the divergence and/or loss of duplicates. This could be realized using public ChIP experiment data, conserved non-coding elements across species [495] or predicted cis-regulatory elements [496]. In parallel with the expression and regulatory evolution it would be interesting to study the pseudogenes and remnants of duplicates, which give insights in the genes which are recently lost (see 6.2.7). The combination of transcriptomic, regulatory and pseudogene data should give a detailed view on how gene loss is happening today long after the WGD event.

Once enough knowledge is gathered it would be possible to accurately model the loss and retention of duplicates and determine the influence of the loss and retention theories. How duplicates are leading to novel functions and how this can be used to explore functions of polyploids in the search of interesting properties. Studying the evolution of genes, genomes and networks can provide valuable knowledge on how species coped with drastic changes through history. This knowledge could help to ensure food production for the increasing population and help to combat the effects of climate change.

CURRICULUM VITAE

Jonas Defoort
20 May 1990
+32 (0) 476 69 35 03
jdefoort@gmail.com
Charles de Kerchovelaan 199A, 9000 Ghent
www.linkedin.com/in/jonasdefoort
http://bioinformatics.psb.ugent.be/beg/people/jofoo

EDUCATION

Master Biochemistry and Biotechnology, Ghent University 2011-2013

- Major in Bioinformatics and Systems Biology
- Minor in Structural Biology
- Master thesis: Integration of molecular interactions in eukaryote gene regulatory networks.

Bachelor Biochemistry and Biotechnology, Ghent University, 2008-2011

ASO Wiskunde-wetenschappen, Spes Nostra Kuurne, 2002-2008

COURSES AND TRAINING FOLLOWED

- Synthetic and Systems Biology summer school (2014)
- Advanced Academic English: Conference Skills (doctoral schools) (2015)
- Speed reading (doctoral schools) (2015)
- EMBO Practical Course on Computational molecular evolution (2016)
- N2N Multidisciplinary Seminar Series on Bioinformatics (doctoral schools) (2016)
- Project Management (doctoral schools) (2016)
- Career guidance (VIB) (2017)

ATTENDANCE OF SYMPOSIA AND CONFERENCES

September 2017| Attendance + Poster presentation

- Plant Genome Evolution 2017, Sitges, Spain

March 2017| Attendance + Poster presentation

- Plant genomics, Amsterdam, The Netherlands

May 2016| Attendance + Poster presentation

- EMBO Practical Course on Computational molecular evolution, Heraklion, Crete

March 2016| Attendance

- ALPHY 2016: French-Belgian meeting on Bioinformatics and Evolutionary Genomics, Lille, France

September 2015| Attendance + Poster presentation

- Plant Genome Evolution 2015, Amsterdam, The Netherlands

June 2014| Attendance + Poster presentation

- Synthetic and Systems Biology 2014, Taormina, Italy

PUBLICATIONS

- Li Z.*, **Defoort J.***, Tasdighian S., Maere S., Van de Peer Y., De Smet R. Gene duplicability of core genes is highly consistent across all angiosperms. The Plant Cell 2016

REFERENCES

1. Van Bel, M., et al., *PLAZA 4.0: an integrative resource for functional, evolutionary and comparative plant genomics.* Nucleic Acids Res, 2017.
2. Weigel, D. and R. Mott, *The 1001 genomes project for Arabidopsis thaliana.* Genome Biol, 2009. **10**(5): p. 107.
3. project, r.g., *The 3,000 rice genomes project.* Gigascience, 2014. **3**: p. 7.
4. Eyre-Walker, A. and P.D. Keightley, *The distribution of fitness effects of new mutations.* Nat Rev Genet, 2007. **8**(8): p. 610-8.
5. Panchy, N., M. Lehti-Shiu, and S.H. Shiu, *Evolution of Gene Duplication in Plants.* Plant Physiol, 2016. **171**(4): p. 2294-316.
6. Sinden, R.R., M.J. Pytlos-Sinden, and V.N. Potaman, *Slipped strand DNA structures.* Front Biosci, 2007. **12**: p. 4788-99.
7. Krsticevic, F.J., et al., *Tandem Duplication Events in the Expansion of the Small Heat Shock Protein Gene Family in Solanum lycopersicum (cv. Heinz 1706).* G3 (Bethesda), 2016. **6**(10): p. 3027-3034.
8. Jiang, N., et al., *Pack-Mutator-like transposable elements (Pack-MULEs) induce directional modification of genes through biased insertion and DNA acquisition.* Proc Natl Acad Sci U S A, 2011. **108**(4): p. 1537-42.
9. Bailey, J.A., et al., *Recent segmental duplications in the human genome.* Science, 2002. **297**(5583): p. 1003-7.
10. Abdelsamad, A. and A. Pecinka, *Pollen-specific activation of Arabidopsis retrogenes is associated with global transcriptional reprogramming.* Plant Cell, 2014. **26**(8): p. 3299-313.
11. Garsmeur, O., et al., *Two evolutionarily distinct classes of paleopolyploidy.* Molecular biology and evolution, 2014. **31**: p. 448-54.
12. Huang, X., et al., *Genomic architecture of heterosis for yield traits in rice.* Nature, 2016. **537**(7622): p. 629-633.
13. Birchler, J.A., D.L. Auger, and N.C. Riddle, *In search of the molecular basis of heterosis.* Plant Cell, 2003. **15**(10): p. 2236-9.
14. Arrigo, N. and M.S. Barker, *Rarely successful polyploids and their legacy in plant genomes.* Curr Opin Plant Biol, 2012. **15**(2): p. 140-6.
15. Mayrose, I., et al., *Methods for studying polyploid diversification and the dead end hypothesis: a reply to Soltis et al. (2014).* New Phytol, 2015. **206**(1): p. 27-35.
16. Van de Peer, Y., E. Mizrachi, and K. Marchal, *The evolutionary significance of polyploidy.* Nat Rev Genet, 2017.
17. Ohno, S., *Evolution by gene duplication*. 1970, New York: Springer. 160.
18. Dehal, P. and J.L. Boore, *Two rounds of whole genome duplication in the ancestral vertebrate.* PLoS Biol, 2005. **3**(10): p. e314.
19. Ruprecht, C., et al., *Revisiting ancestral polyploidy in plants.* Sci Adv, 2017. **3**(7): p. e1603195.
20. Bowers, J.E., et al., *Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events.* Nature, 2003. **422**(6930): p. 433-8.
21. Vanneste, K., et al., *Analysis of 41 plant genomes supports a wave of successful genome duplications in association with the Cretaceous-Paleogene boundary.* Genome research, 2014.
22. Vanneste, K., S. Maere, and Y. Van de Peer, *Tangled up in two: a burst of genome duplications at the end of the Cretaceous and the consequences for plant evolution.* Philosophical Transactions of the Royal Society B-Biological Sciences, 2014. **369**(1648).
23. Van de Peer, Y., S. Maere, and A. Meyer, *The evolutionary significance of ancient genome duplications.* Nature Reviews Genetics, 2009. **10**: p. 725-732.
24. Fawcett, J.a., S. Maere, and Y. Van de Peer, *Plants with double genomes might have had a better chance to survive the Cretaceous-Tertiary extinction event.* Proceedings of the National Academy of Sciences of the United States of America, 2009. **106**: p. 5737-42.
25. Novikova, P.Y., N. Hohmann, and Y. Van de Peer, *Polyploid Arabidopsis species originated around recent glaciation maxima.* Curr Opin Plant Biol, 2018. **42**: p. 8-15.

26. Van de Peer, Y., *Computational approaches to unveiling ancient genome duplications.* Nat Rev Genet, 2004. **5**(10): p. 752-63.
27. Vanneste, K., Y. Van de Peer, and S. Maere, *Inference of genome duplications from age distributions revisited.* Mol Biol Evol, 2013. **30**(1): p. 177-90.
28. Lynch, M. and J.S. Conery, *The evolutionary fate and consequences of duplicate genes.* Science, 2000. **290**(5494): p. 1151-5.
29. Maere, S., et al., *Modeling gene and genome duplications in eukaryotes.* Proc Natl Acad Sci U S A, 2005. **102**(15): p. 5454-9.
30. Durand, D. and R. Hoberman, *Diagnosing duplications--can it be done?* Trends Genet, 2006. **22**(3): p. 156-64.
31. Bevan, M.W., et al., *Genomic innovation for crop improvement.* Nature, 2017. **543**(7645): p. 346-354.
32. Schnable, J.C., N.M. Springer, and M. Freeling, *Differentiation of the maize subgenomes by genome dominance and both ancient and ongoing gene loss.* Proceedings of the National Academy of Sciences of the United States of America, 2011. **108**: p. 4069-74.
33. De Smet, R., et al., *Convergent gene loss following gene and genome duplications creates single-copy families in flowering plants.* Proc Natl Acad Sci U S A, 2013. **110**(8): p. 2898-903.
34. De Smet, R. and Y. Van de Peer, *Redundancy and rewiring of genetic networks following genome-wide duplication events.* Current opinion in plant biology, 2012. **15**: p. 168-76.
35. Gout, J.-f., M. Lynch, and J.-f. Gout, *Maintenance and loss of duplicated genes by dosage subfunctionalization.* 2015.
36. Birchler, J.A. and R.A. Veitia, *The gene balance hypothesis: implications for gene regulation, quantitative traits and evolution.* The New phytologist, 2010. **186**: p. 54-62.
37. Birchler, J.a., et al., *Dosage balance in gene regulation: biological implications.* Trends in genetics : TIG, 2005. **21**: p. 219-26.
38. Veitia, R.a., S. Bottani, and J.a. Birchler, *Gene dosage effects: nonlinearities, genetic interactions, and dosage compensation.* Trends in genetics : TIG, 2013. **29**: p. 385-93.
39. Veitia, R.a., *Nonlinear effects in macromolecular assembly and dosage sensitivity.* Journal of theoretical biology, 2003. **220**: p. 19-25.
40. Veitia, R.a., *On gene dosage balance in protein complexes: a comment on Semple JI, Vavouri T, Lehner B. A simple principle concerning the robustness of protein complex activity to changes in gene expression.* BMC systems biology, 2009. **3**: p. 16.
41. Veitia, R.A., *A generalized model of gene dosage and dominant negative effects in macromolecular complexes.* FASEB journal : official publication of the Federation of American Societies for Experimental Biology, 2010. **24**: p. 994-1002.
42. Hudson, C.M., et al., *Selection for higher gene copy number after different types of plant gene duplications.* Genome Biol Evol, 2011. **3**: p. 1369-80.
43. Kaltenegger, E. and D. Ober, *Paralogue Interference Affects the Dynamics after Gene Duplication.* Trends in Plant Science, 2015. **20**: p. 814-821.
44. Des Marais, D.L. and M.D. Rausher, *Escape from adaptive conflict after duplication in an anthocyanin pathway gene.* Nature, 2008. **454**(7205): p. 762-5.
45. Voordeckers, K. and K.J. Verstrepen, *Experimental evolution of the model eukaryote Saccharomyces cerevisiae yields insight into the molecular mechanisms underlying adaptation.* Curr Opin Microbiol, 2015. **28**: p. 1-9.
46. Albalat, R. and C. Canestro, *Evolution by gene loss.* Nature Reviews Genetics, 2016. **17**(7): p. 379-391.
47. Donoghue, M.T., et al., *Evolutionary origins of Brassicaceae specific genes in Arabidopsis thaliana.* BMC Evol Biol, 2011. **11**: p. 47.
48. Conant, G.C., *Comparative Genomics as a Time Machine: How Relative Gene Dosage and Metabolic Requirements Shaped the Time-dependent Resolution of Yeast Polyploidy.* Molecular Biology and Evolution, 2014. **31**: p. 3184-3193.
49. Blanc, G. and K.H. Wolfe, *Functional divergence of duplicated genes formed by polyploidy during Arabidopsis evolution.* The Plant cell, 2004. **16**: p. 1679-91.
50. Maclean, C.J. and D. Greig, *Reciprocal gene loss following experimental whole-genome duplication causes reproductive isolation in yeast.* Evolution, 2011. **65**(4): p. 932-45.

51. McGrath, C.L., et al., *Differential retention and divergent resolution of duplicate genes following whole-genome duplication.* Genome Res, 2014. **24**(10): p. 1665-75.
52. Bekaert, M., et al., *Two-phase resolution of polyploidy in the Arabidopsis metabolic network gives rise to relative and absolute dosage constraints.* The Plant cell, 2011. **23**: p. 1719-28.
53. Emmert-Streib, F., M. Dehmer, and B. Haibe-Kains, *Gene regulatory networks and their applications: understanding biological and medical problems in terms of networks.* Front Cell Dev Biol, 2014. **2**: p. 38.
54. Jaenisch, R. and A. Bird, *Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals.* Nat Genet, 2003. **33 Suppl**: p. 245-54.
55. Glisovic, T., et al., *RNA-binding proteins and post-transcriptional gene regulation.* FEBS Lett, 2008. **582**(14): p. 1977-86.
56. Lothrop, A.P., M.P. Torres, and S.M. Fuchs, *Deciphering post-translational modification codes.* FEBS Lett, 2013. **587**(8): p. 1247-57.
57. Hudson, W.H. and E.A. Ortlund, *The structure, function and evolution of proteins that bind DNA and RNA.* Nat Rev Mol Cell Biol, 2014. **15**(11): p. 749-60.
58. Grosshans, H. and W. Filipowicz, *Molecular biology: the expanding world of small RNAs.* Nature, 2008. **451**(7177): p. 414-6.
59. Clark, M.B., et al., *The dark matter rises: the expanding world of regulatory RNAs.* Essays Biochem, 2013. **54**: p. 1-16.
60. Keskin, O., N. Tuncbag, and A. Gursoy, *Predicting Protein-Protein Interactions from the Molecular to the Proteome Level.* Chem Rev, 2016. **116**(8): p. 4884-909.
61. Van Landeghem, S., et al., *The potential of text mining in data integration and network biology for plant research: a case study on Arabidopsis.* The Plant cell, 2013. **25**: p. 794-807.
62. Franceschini, A., et al., *STRING v9.1: protein-protein interaction networks, with increased coverage and integration.* Nucleic acids research, 2013. **41**: p. D808-15.
63. Consortium, A.I.M., *Evidence for network evolution in an Arabidopsis interactome map.* Science (New York, N.Y.), 2011. **333**: p. 601-7.
64. Consortium, E.P., *The ENCODE (ENCyclopedia Of DNA Elements) Project.* Science, 2004. **306**(5696): p. 636-40.
65. mod, E.C., et al., *Identification of functional elements and regulatory circuits by Drosophila modENCODE.* Science, 2010. **330**(6012): p. 1787-97.
66. Rao, V.S., et al., *Protein-protein interaction detection: methods and analysis.* Int J Proteomics, 2014. **2014**: p. 147648.
67. Braun, P. and A.C. Gingras, *History of protein-protein interactions: from egg-white to complex networks.* Proteomics, 2012. **12**(10): p. 1478-98.
68. Dey, B., et al., *DNA-protein interactions: methods for detection and analysis.* Mol Cell Biochem, 2012. **365**(1-2): p. 279-99.
69. Si, J., R. Zhao, and R. Wu, *An overview of the prediction of protein DNA-binding sites.* Int J Mol Sci, 2015. **16**(3): p. 5194-215.
70. Oulas, A., et al., *Prediction of miRNA targets.* Methods Mol Biol, 2015. **1269**: p. 207-29.
71. Anders, G., et al., *doRiNA: a database of RNA interactions in post-transcriptional regulation.* Nucleic acids research, 2012. **40**: p. D180-6.
72. Mani, R., et al., *Defining genetic interaction.* Proc Natl Acad Sci U S A, 2008. **105**(9): p. 3461-6.
73. Koch, E.N., et al., *Conserved rules govern genetic interaction degree across species.* Genome biology, 2012. **13**: p. R57.
74. Ma, X. and L. Gao, *Biological network analysis: insights into structure and functions.* Brief Funct Genomics, 2012. **11**(6): p. 434-42.
75. Zotenko, E., et al., *Why do hubs in the yeast protein interaction network tend to be essential: reexamining the connection between the network topology and essentiality.* PLoS computational biology, 2008. **4**: p. e1000140.
76. Shoval, O. and U. Alon, *SnapShot: network motifs.* Cell, 2010. **143**: p. 326-e1.
77. Alon, U., *Network motifs: theory and experimental approaches.* Nature reviews. Genetics, 2007. **8**: p. 450-61.

78. Yeger-Lotem, E., et al., *Network motifs in integrated cellular networks of transcription-regulation and protein-protein interaction.* Proceedings of the National Academy of Sciences of the United States of America, 2004. **101**: p. 5934-9.
79. Zhang, Y., et al., *Network motif-based identification of transcription factor-target gene relationships by integrating multi-source biological data.* BMC bioinformatics, 2008. **9**: p. 203.
80. Zhang, L.V., et al., *Motifs, themes and thematic maps of an integrated Saccharomyces cerevisiae interaction network.* 2005.
81. Martinez, N.J., et al., *A C. elegans genome-scale microRNA network contains composite feedback motifs with high flux capacity.* Genes & development, 2008. **22**: p. 2535-49.
82. Yu, D., et al., *Review of biological network data and its applications.* Genomics Inform, 2013. **11**(4): p. 200-10.
83. Narang, V., et al., *Automated Identification of Core Regulatory Genes in Human Gene Regulatory Networks.* PLoS Comput Biol, 2015. **11**(9): p. e1004504.
84. Park, J., et al., *Identifying functional gene regulatory network phenotypes underlying single cell transcriptional variability.* Prog Biophys Mol Biol, 2015. **117**(1): p. 87-98.
85. Lockhart, D.J., et al., *Expression monitoring by hybridization to high-density oligonucleotide arrays.* Nat Biotechnol, 1996. **14**(13): p. 1675-80.
86. Heber, S. and B. Sick, *Quality assessment of Affymetrix GeneChip data.* OMICS, 2006. **10**(3): p. 358-68.
87. Irizarry, R.A., et al., *Summaries of Affymetrix GeneChip probe level data.* Nucleic Acids Res, 2003. **31**(4): p. e15.
88. Wang, Z., M. Gerstein, and M. Snyder, *RNA-Seq: a revolutionary tool for transcriptomics.* Nat Rev Genet, 2009. **10**(1): p. 57-63.
89. Conesa, A., et al., *A survey of best practices for RNA-seq data analysis.* Genome Biol, 2016. **17**: p. 13.
90. Liang, J., W. Cai, and Z. Sun, *Single-cell sequencing technologies: current and future.* J Genet Genomics, 2014. **41**(10): p. 513-28.
91. Domon, B. and R. Aebersold, *Mass spectrometry and protein analysis.* Science, 2006. **312**(5771): p. 212-7.
92. Lo, C.A., et al., *Quantification of Protein Levels in Single Living Cells.* Cell Rep, 2015. **13**(11): p. 2634-2644.
93. Gene Ontology, C., *Gene Ontology Consortium: going forward.* Nucleic Acids Res, 2015. **43**(Database issue): p. D1049-56.
94. Maere, S., K. Heymans, and M. Kuiper, *BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks.* Bioinformatics (Oxford, England), 2005. **21**: p. 3448-9.
95. Woods, S., et al., *Duplication and retention biases of essential and non-essential genes revealed by systematic knockdown analyses.* PLoS Genet, 2013. **9**(5): p. e1003330.
96. Davis, J.C. and D.A. Petrov, *Preferential duplication of conserved proteins in eukaryotic genomes.* PLoS Biol, 2004. **2**(3): p. E55.
97. Koonin, E.V., et al., *A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes.* Genome Biol, 2004. **5**(2): p. R7.
98. Liang, H., et al., *Protein under-wrapping causes dosage sensitivity and decreases gene duplicability.* PLoS Genet, 2008. **4**(1): p. e11.
99. Rambaldi, D., et al., *Low duplicability and network fragility of cancer genes.* Trends in Genetics, 2008. **24**(9): p. 427-430.
100. Makino, T., K. Hokamp, and A. McLysaght, *The complex relationship of gene duplication and essentiality.* Trends in Genetics, 2009. **25**(4): p. 152-155.
101. He, X.L. and J.Z. Zhang, *Higher duplicability of less important genes in yeast genomes.* Molecular Biology and Evolution, 2006. **23**(1): p. 144-151.
102. Aury, J.M., et al., *Global trends of whole-genome duplications revealed by the ciliate Paramecium tetraurelia.* Nature, 2006. **444**(7116): p. 171-8.
103. Papp, B., C. Pal, and L.D. Hurst, *Dosage sensitivity and the evolution of gene families in yeast.* Nature, 2003. **424**(6945): p. 194-7.

104. Freeling, M., *Bias in plant gene content following different sorts of duplication: tandem, whole-genome, segmental, or by transposition.* Annu Rev Plant Biol, 2009. **60**: p. 433-53.
105. Blanc, G. and K.H. Wolfe, *Functional divergence of duplicated genes formed by polyploidy during Arabidopsis evolution.* Plant Cell, 2004. **16**(7): p. 1679-91.
106. Seoighe, C. and C. Gehring, *Genome duplication led to highly selective expansion of the Arabidopsis thaliana proteome.* Trends Genet, 2004. **20**(10): p. 461-4.
107. Van de Peer, Y., et al., *The flowering world: a tale of duplications.* Trends Plant Sci, 2009. **14**(12): p. 680-8.
108. Vanneste, K., et al., *Analysis of 41 plant genomes supports a wave of successful genome duplications in association with the Cretaceous-Paleogene boundary.* Genome Res, 2014. **24**(8): p. 1334-47.
109. Birchler, J.A., et al., *Dosage balance in gene regulation: biological implications.* Trends Genet, 2005. **21**(4): p. 219-26.
110. Birchler, J.A. and R.A. Veitia, *The gene balance hypothesis: from classical genetics to modern genomics.* Plant Cell, 2007. **19**(2): p. 395-402.
111. Birchler, J.A. and R.A. Veitia, *Gene balance hypothesis: Connecting issues of dosage sensitivity across biological disciplines.* Proc Natl Acad Sci U S A, 2012. **109**(37): p. 14746-53.
112. Edger, P.P. and J.C. Pires, *Gene and genome duplications: the impact of dosage-sensitivity on the fate of nuclear genes.* Chromosome Res, 2009. **17**(5): p. 699-717.
113. Rodgers-Melnick, E., et al., *Contrasting patterns of evolution following whole genome versus tandem duplication events in Populus.* Genome Res, 2012. **22**(1): p. 95-105.
114. Makino, T. and A. McLysaght, *Ohnologs in the human genome are dosage balanced and frequently associated with disease.* Proc Natl Acad Sci U S A, 2010. **107**(20): p. 9270-4.
115. Brunet, F.G., et al., *Gene loss and evolutionary rates following whole-genome duplication in teleost fishes.* Molecular Biology and Evolution, 2006. **23**(9): p. 1808-1816.
116. Huminiecki, L. and C.H. Heldin, *2R and remodeling of vertebrate signal transduction engine.* BMC Biol, 2010. **8**: p. 146.
117. Schmutz, J., et al., *Genome sequence of the palaeopolyploid soybean.* Nature, 2010. **463**(7278): p. 178-83.
118. Douglas, G.M., et al., *Hybrid origins and the earliest stages of diploidization in the highly successful recent polyploid Capsella bursa-pastoris.* Proc Natl Acad Sci U S A, 2015. **112**(9): p. 2806-11.
119. Buggs, R.J., et al., *Rapid, repeated, and clustered loss of duplicate genes in allopolyploid plant populations of independent origin.* Curr Biol, 2012. **22**(3): p. 248-52.
120. McGrath, C.L. and M. Lynch, *Evolutionary Significance of Whole-Genome Duplication*, in *Polyploidy and Genome Evolution*, P.S. Soltis and D.E. Soltis, Editors. 2012, Springer-Verlag: Berlin. p. 1-20.
121. Duarte, J.M., et al., *Identification of shared single copy nuclear genes in Arabidopsis, Populus, Vitis and Oryza and their phylogenetic utility across various taxonomic levels.* BMC Evol Biol, 2010. **10**: p. 61.
122. Armisen, D., A. Lecharny, and S. Aubourg, *Unique genes in plants: specificities and conserved features throughout evolution.* BMC Evol Biol, 2008. **8**: p. 280.
123. Han, F.M., et al., *Identification, characterization, and utilization of single copy genes in 29 angiosperm genomes.* Bmc Genomics, 2014. **15**.
124. Paterson, A.H., et al., *Many gene and domain families have convergent fates following independent whole-genome duplication events in Arabidopsis, Oryza, Saccharomyces and Tetraodon.* Trends Genet, 2006. **22**(11): p. 597-602.
125. Soltis, D.E., et al., *What we still don't know about polyploidy.* Taxon, 2010. **59**(5): p. 1387-1403.
126. Barker, M.S., et al., *Multiple Paleopolyploidizations during the Evolution of the Compositae Reveal Parallel Patterns of Duplicate Gene Retention after Millions of Years.* Molecular Biology and Evolution, 2008. **25**(11): p. 2445-2455.
127. Carretero-Paulet, L. and M.A. Fares, *Evolutionary dynamics and functional specialization of plant paralogs formed by whole and small-scale genome duplications.* Mol Biol Evol, 2012. **29**(11): p. 3541-51.

128. Conant, G.C., *Comparative Genomics as a Time Machine: How Relative Gene Dosage and Metabolic Requirements Shaped the Time-dependent Resolution of Yeast Polyploidy.* Molecular Biology and Evolution, 2014. **31**(12): p. 3184-3193.
129. Jiao, Y., et al., *Integrated syntenic and phylogenomic analyses reveal an ancient genome duplication in monocots.* Plant Cell, 2014. **26**(7): p. 2792-802.
130. Bailey, N., *The elements of stochastic processes*. 1964, New York: John Wiley & Sons, Inc.
131. Rabier, C.E., T. Ta, and C. Ane, *Detecting and locating whole genome duplications on a phylogeny: a probabilistic approach.* Mol Biol Evol, 2014. **31**(3): p. 750-62.
132. Lloyd, A.H., et al., *Meiotic gene evolution: can you teach a new dog new tricks?* Mol Biol Evol, 2014. **31**(7): p. 1724-7.
133. Vanneste, K., Y. Van de Peer, and S. Maere, *Inference of Genome Duplications from Age Distributions Revisited.* Mol Biol Evol, 2012. **30**(1): p. 177-190.
134. Blanc, G. and K.H. Wolfe, *Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes.* Plant Cell, 2004. **16**(7): p. 1667-78.
135. Lynch, M. and J.S. Conery, *The evolutionary demography of duplicate genes.* J Struct Funct Genomics, 2003. **3**(1-4): p. 35-44.
136. Monti, S., et al., *Consensus clustering: A resampling-based method for class discovery and visualization of gene expression microarray data.* Machine Learning, 2003. **52**(1-2): p. 91-118.
137. Amborella Genome Project, *The Amborella genome and the evolution of flowering plants.* Science, 2013. **342**(6165): p. 1241089.
138. Smith, S.A. and M.J. Donoghue, *Rates of molecular evolution are linked to life history in flowering plants.* Science, 2008. **322**(5898): p. 86-9.
139. Myouga, F., et al., *The Chloroplast Function Database II: A Comprehensive Collection of Homozygous Mutants and Their Phenotypic/Genotypic Traits for Nuclear-Encoded Chloroplast Proteins.* Plant and Cell Physiology, 2013. **54**(2): p. E2-+.
140. Perez-Rodriguez, P., et al., *PlnTFDB: updated content and new features of the plant transcription factor database.* Nucleic Acids Res, 2010. **38**(Database issue): p. D822-7.
141. Freeling, M. and B.C. Thomas, *Gene-balanced duplications, like tetraploidy, provide predictable drive to increase morphological complexity.* Genome Res, 2006. **16**(7): p. 805-14.
142. Lloyd, J. and D. Meinke, *A comprehensive dataset of genes with a loss-of-function mutant phenotype in Arabidopsis.* Plant Physiol, 2012. **158**(3): p. 1115-29.
143. Lloyd, J.P., et al., *Characteristics of Plant Essential Genes Allow for within-and between-Species Prediction of Lethal Mutant Phenotypes.* Plant Cell, 2015. **27**(8): p. 2133-2147.
144. Conant, G.C., J.A. Birchler, and J.C. Pires, *Dosage, duplication, and diploidization: clarifying the interplay of multiple models for duplicate gene evolution over time.* Curr Opin Plant Biol, 2014. **19**: p. 91-8.
145. Wang, Y.P., et al., *Modes of Gene Duplication Contribute Differently to Genetic Novelty and Redundancy, but Show Parallels across Divergent Angiosperms.* Plos One, 2011. **6**(12).
146. Woodhouse, M.R., H.B. Tang, and M. Freeling, *Different Gene Families in Arabidopsis thaliana Transposed in Different Epochs and at Different Frequencies throughout the Rosids.* Plant Cell, 2011. **23**(12): p. 4241-4253.
147. Freeling, M., et al., *Many or most genes in Arabidopsis transposed after the origin of the order Brassicales.* Genome Research, 2008. **18**(12): p. 1924-1937.
148. Wapinski, I., et al., *Natural history and evolutionary principles of gene duplication in fungi.* Nature, 2007. **449**(7158): p. 54-61.
149. Molina, N. and E. van Nimwegen, *Scaling laws in functional genome content across prokaryotic clades and lifestyles.* Trends in Genetics, 2009. **25**(6): p. 243-247.
150. van Nimwegen, E., *Scaling laws in the functional content of genomes.* Trends in Genetics, 2003. **19**(9): p. 479-484.
151. Kitano, H., *Biological robustness.* Nat Rev Genet, 2004. **5**(11): p. 826-37.
152. Siegel, J.J. and A. Amon, *New Insights into the Troubles of Aneuploidy.* Annual Review of Cell and Developmental Biology, Vol 28, 2012. **28**: p. 189-214.

153. Dean, E.J., et al., *Pervasive and Persistent Redundancy among Duplicated Genes in Yeast.* Plos Genetics, 2008. **4**(7).
154. Bridgham, J.T., et al., *Evolution of a new function by degenerative mutation in cephalochordate steroid receptors.* PLoS Genet, 2008. **4**(9): p. e1000191.
155. Kaltenegger, E. and D. Ober, *Paralogue Interference Affects the Dynamics after Gene Duplication.* Trends in Plant Science, 2015: p. In Press.
156. Zhang, J. and J.R. Yang, *Determinants of the rate of protein sequence evolution.* Nat Rev Genet, 2015. **16**(7): p. 409-20.
157. Lynch, M. and V. Katju, *The altered evolutionary trajectories of gene duplicates.* Trends Genet, 2004. **20**(11): p. 544-9.
158. Bekaert, M., et al., *Two-phase resolution of polyploidy in the Arabidopsis metabolic network gives rise to relative and absolute dosage constraints.* Plant Cell, 2011. **23**(5): p. 1719-28.
159. Yona, A.H., et al., *Chromosomal duplication is a transient evolutionary solution to stress.* Proceedings of the National Academy of Sciences of the United States of America, 2012. **109**(51): p. 21010-21015.
160. Sunshine, A.B., et al., *The fitness consequences of aneuploidy are driven by condition-dependent gene effects.* PLoS Biol, 2015. **13**(5): p. e1002155.
161. te Beest, M., et al., *The more the better? The role of polyploidy in facilitating plant invasions.* Ann Bot, 2012. **109**(1): p. 19-45.
162. Selmecki, A.M., et al., *Polyploidy can drive rapid adaptation in yeast.* Nature, 2015. **519**(7543): p. 349-52.
163. Chao, D.Y., et al., *Polyploids exhibit higher potassium uptake and salinity tolerance in Arabidopsis.* Science, 2013. **341**(6146): p. 658-9.
164. Van de Peer, Y., S. Maere, and A. Meyer, *The evolutionary significance of ancient genome duplications.* Nat Rev Genet, 2009. **10**(10): p. 725-32.
165. Hahn, M.A., M. van Kleunen, and H. Muller-Scharer, *Increased Phenotypic Plasticity to Climate May Have Boosted the Invasion Success of Polyploid Centaurea stoebe.* Plos One, 2012. **7**(11).
166. Yang, C., et al., *Evolution of physiological responses to salt stress in hexaploid wheat.* Proc Natl Acad Sci U S A, 2014. **111**(32): p. 11882-7.
167. Gresham, D., et al., *The repertoire and dynamics of evolutionary adaptations to controlled nutrient-limited environments in yeast.* PLoS Genet, 2008. **4**(12): p. e1000303.
168. Dunham, M.J., et al., *Characteristic genome rearrangements in experimental evolution of Saccharomyces cerevisiae.* Proc Natl Acad Sci U S A, 2002. **99**(25): p. 16144-9.
169. Selmecki, A., A. Forche, and J. Berman, *Aneuploidy and isochromosome formation in drug-resistant Candida albicans.* Science, 2006. **313**(5785): p. 367-70.
170. Drummond, D.A. and C.O. Wilke, *Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution.* Cell, 2008. **134**(2): p. 341-352.
171. Alvarez-Ponce, D. and M.A. Fares, *Evolutionary rate and duplicability in the Arabidopsis thaliana protein-protein interaction network.* Genome Biol Evol, 2012. **4**(12): p. 1263-74.
172. Chae, L., et al., *Towards understanding how molecular networks evolve in plants.* Current Opinion in Plant Biology, 2012. **15**(2): p. 177-184.
173. D'Antonio, M. and F.D. Ciccarelli, *Modification of gene duplicability during the evolution of protein interaction network.* PLoS Comput Biol, 2011. **7**(4): p. e1002029.
174. Slotte, T., et al., *The Capsella rubella genome and the genomic consequences of rapid mating system evolution.* Nat Genet, 2013. **45**(7): p. 831-5.
175. Li, L., C.J. Stoeckert, Jr., and D.S. Roos, *OrthoMCL: identification of ortholog groups for eukaryotic genomes.* Genome Res, 2003. **13**(9): p. 2178-89.
176. Edgar, R.C., *MUSCLE: multiple sequence alignment with high accuracy and high throughput.* Nucleic Acids Res, 2004. **32**(5): p. 1792-7.
177. Capella-Gutierrez, S., J.M. Silla-Martinez, and T. Gabaldon, *trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses.* Bioinformatics, 2009. **25**(15): p. 1972-3.
178. Gil, M., et al., *CodonPhyML: fast maximum likelihood phylogeny estimation under codon substitution models.* Mol Biol Evol, 2013. **30**(6): p. 1270-80.

179. Muse, S.V. and B.S. Gaut, *A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome.* Mol Biol Evol, 1994. **11**(5): p. 715-24.
180. Goldman, N. and Z. Yang, *A codon-based model of nucleotide substitution for protein-coding DNA sequences.* Mol Biol Evol, 1994. **11**(5): p. 725-36.
181. Yap, V.B., et al., *Estimates of the effect of natural selection on protein-coding content.* Mol Biol Evol, 2010. **27**(3): p. 726-34.
182. Stamatakis, A., *RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies.* Bioinformatics, 2014. **30**(9): p. 1312-3.
183. Salichos, L. and A. Rokas, *Inferring ancient divergences requires genes with strong phylogenetic signals.* Nature, 2013. **497**(7449): p. 327-31.
184. Salichos, L., A. Stamatakis, and A. Rokas, *Novel Information Theory-Based Measures for Quantifying Incongruence among Phylogenetic Trees.* Molecular Biology and Evolution, 2014. **31**(5): p. 1261-1271.
185. Bremer, B., et al., *An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG III.* Botanical Journal of the Linnean Society, 2009. **161**(2): p. 105-121.
186. Guindon, S., et al., *New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0.* Systematic Biology, 2010. **59**(3): p. 307-321.
187. Anisimova, M. and O. Gascuel, *Approximate likelihood-ratio test for branches: A fast, accurate, and powerful alternative.* Syst Biol, 2006. **55**(4): p. 539-52.
188. Stolzer, M., et al., *Inferring duplications, losses, transfers and incomplete lineage sorting with nonbinary species trees.* Bioinformatics, 2012. **28**(18): p. i409-i415.
189. Hahn, M.W., *Bias in phylogenetic tree reconciliation methods: implications for vertebrate genome evolution.* Genome Biol, 2007. **8**(7): p. R141.
190. Wu, Y.C., et al., *TreeFix: Statistically Informed Gene Tree Error Correction Using Species Trees.* Systematic Biology, 2013. **62**(1): p. 110-120.
191. Nguyen, T.H., et al., *Support measures to estimate the reliability of evolutionary events predicted by reconciliation methods.* PLoS One, 2013. **8**(10): p. e73667.
192. Oliver, T., et al., *Using reconfigurable hardware to accelerate multiple sequence alignment with ClustalW.* Bioinformatics, 2005. **21**(16): p. 3431-2.
193. Hall, B.G., *Phylogenetic trees made easy*. 2004: Sinauer Associates.
194. Suyama, M., D. Torrents, and P. Bork, *PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments.* Nucleic Acids Research, 2006. **34**: p. W609-W612.
195. Yang, Z.H., *PAML: a program package for phylogenetic analysis by maximum likelihood.* Computer Applications in the Biosciences, 1997. **13**(5): p. 555-556.
196. Yang, Z.H., *PAML 4: Phylogenetic analysis by maximum likelihood.* Molecular Biology and Evolution, 2007. **24**(8): p. 1586-1591.
197. Yang, Z. and R. Nielsen, *Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models.* Mol Biol Evol, 2000. **17**(1): p. 32-43.
198. Nei, M. and T. Gojobori, *Simple Methods for Estimating the Numbers of Synonymous and Nonsynonymous Nucleotide Substitutions.* Molecular Biology and Evolution, 1986. **3**(5): p. 418-426.
199. Hahn, M.W., et al., *Estimating the tempo and mode of gene family evolution from comparative genomic data.* Genome Res, 2005. **15**(8): p. 1153-60.
200. Wilkerson, M.D. and D.N. Hayes, *ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking.* Bioinformatics, 2010. **26**(12): p. 1572-3.
201. Chatr-Aryamontri, A., et al., *The BioGRID interaction database: 2013 update.* Nucleic Acids Res, 2013. **41**(Database issue): p. D816-23.
202. De Bodt, S., et al., *CORNET 2.0: integrating plant coexpression, protein-protein interactions, regulatory interactions, gene associations and functional annotations.* New Phytol, 2012. **195**(3): p. 707-20.
203. Franceschini, A., et al., *STRING v9.1: protein-protein interaction networks, with increased coverage and integration.* Nucleic Acids Res, 2013. **41**(Database issue): p. D808-15.

204. Van Landeghem, S., et al., *Large-scale event extraction from literature with multi-level gene normalization.* PLoS One, 2013. **8**(4): p. e55814.
205. Di Rubbo, S., et al., *The clathrin adaptor complex AP-2 mediates endocytosis of brassinosteroid insensitive1 in Arabidopsis.* Plant Cell, 2013. **25**(8): p. 2986-97.
206. Van Leene, J., et al., *Targeted interactomics reveals a complex core cell cycle machinery in Arabidopsis thaliana.* Mol Syst Biol, 2010. **6**: p. 397.
207. Gadeyne, A., et al., *The TPLATE adaptor complex drives clathrin-mediated endocytosis in plants.* Cell, 2014. **156**(4): p. 691-704.
208. Bassard, J.E., et al., *Protein-protein and protein-membrane associations in the lignin pathway.* Plant Cell, 2012. **24**(11): p. 4465-82.
209. Domenichini, S., et al., *Evidence for a role of Arabidopsis CDT1 proteins in gametophyte development and maintenance of genome integrity.* Plant Cell, 2012. **24**(7): p. 2779-91.
210. Takahashi, N., et al., *The DNA replication checkpoint aids survival of plants deficient in the novel replisome factor ETG1.* EMBO J, 2008. **27**(13): p. 1840-51.
211. Pauwels, L., et al., *NINJA connects the co-repressor TOPLESS to jasmonate signalling.* Nature, 2010. **464**(7289): p. 788-91.
212. Fonseca, S., et al., *bHLH003, bHLH013 and bHLH017 are new targets of JAZ repressors negatively regulating JA responses.* PLoS One, 2014. **9**(1): p. e86182.
213. Cromer, L., et al., *Centromeric Cohesion Is Protected Twice at Meiosis, by SHUGOSHINs at Anaphase 1 and by PATRONUS at Interkinesis.* Current Biology, 2013. **23**(21): p. 2090-2099.
214. Antoni, R., et al., *PYRABACTIN RESISTANCE1-LIKE8 plays an important role for the regulation of abscisic acid signaling in root.* Plant Physiol, 2013. **161**(2): p. 931-41.
215. Eloy, N.B., et al., *SAMBA, a plant-specific anaphase-promoting complex/cyclosome regulator is involved in early development and A-type cyclin stabilization.* Proc Natl Acad Sci U S A, 2012. **109**(34): p. 13853-8.
216. Perez, A.C., et al., *The Non-JAZ TIFY Protein TIFY8 from Arabidopsis thaliana Is a Transcriptional Repressor.* Plos One, 2014. **9**(1).
217. Spinner, L., et al., *A protein phosphatase 2A complex spatially controls plant cell division.* Nat Commun, 2013. **4**: p. 1863.
218. Heijde, M., et al., *Constitutively active UVR8 photoreceptor variant in Arabidopsis.* Proc Natl Acad Sci U S A, 2013. **110**(50): p. 20326-31.
219. Vercruyssen, L., et al., *ANGUSTIFOLIA3 binds to SWI/SNF chromatin remodeling complexes to regulate transcription during Arabidopsis leaf development.* Plant Cell, 2014. **26**(1): p. 210-29.
220. Maere, S., K. Heymans, and M. Kuiper, *BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks.* Bioinformatics, 2005. **21**(16): p. 3448-9.
221. Benjamini, Y. and Y. Hochberg, *Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing.* Journal of the Royal Statistical Society Series B-Methodological, 1995. **57**(1): p. 289-300.
222. Conant, G.C. and K.H. Wolfe, *Turning a hobby into a job: how duplicated genes find new functions.* Nat Rev Genet, 2008. **9**(12): p. 938-50.
223. Soltis, D.E., et al., *Polyploidy and angiosperm diversification.* Am J Bot, 2009. **96**(1): p. 336-48.
224. Jiao, Y., et al., *Ancestral polyploidy in seed plants and angiosperms.* Nature, 2011. **473**(7345): p. 97-100.
225. Jiao, Y., et al., *A genome triplication associated with early diversification of the core eudicots.* Genome Biol, 2012. **13**(1): p. R3.
226. Casneuf, T., et al., *Nonrandom divergence of gene expression following gene and genome duplications in the flowering plant Arabidopsis thaliana.* Genome biology, 2006. **7**: p. R13.
227. Wang, Y., et al., *Modes of gene duplication contribute differently to genetic novelty and redundancy, but show parallels across divergent angiosperms.* PLoS ONE, 2011. **6**.
228. Carretero-Paulet, L. and M.A. Fares, *Evolutionary dynamics and functional specialization of plant paralogs formed by whole and small-scale genome duplications.* Molecular biology and evolution, 2012. **29**: p. 3541-51.

229. Li, Z., et al., *Gene Duplicability of Core Genes Is Highly Consistent across All Angiosperms.* Plant Cell, 2016. **28**(2): p. 326-44.
230. Wapinski, I., et al., *Natural history and evolutionary principles of gene duplication in fungi.* Nature, 2007. **449**: p. 54-61.
231. Makino, T. and A. McLysaght, *Positionally biased gene loss after whole genome duplication: evidence from human, yeast, and plant.* Genome research, 2012. **22**: p. 2427-35.
232. Fares, M.A., et al., *The Roles of Whole-Genome and Small-Scale Duplications in the Functional Specialization of Saccharomyces cerevisiae Genes.* PLoS Genetics, 2013. **9**.
233. Makino, T., K. Hokamp, and A. McLysaght, *The complex relationship of gene duplication and essentiality.* Trends in Genetics, 2009. **25**: p. 152-155.
234. Alvarez-Ponce, D. and M.A. Fares, *Evolutionary rate and duplicability in the Arabidopsis thaliana protein-protein interaction network.* Genome Biology and Evolution, 2012. **4**: p. 1263-1274.
235. Guo, H., et al., *Function Relaxation Followed by Diversifying Selection after Whole-Genome Duplication in Flowering Plants.* Plant Physiology, 2013. **162**: p. 769-778.
236. Freeling, M. and B.C. Thomas, *Gene-balanced duplications, like tetraploidy, provide predictable drive to increase morphological complexity.* Genome research, 2006. **16**: p. 805-14.
237. Makino, T. and A. McLysaght, *Ohnologs in the human genome are dosage balanced and frequently associated with disease.* Proceedings of the National Academy of Sciences of the United States of America, 2010. **107**: p. 9270-4.
238. Tasdighian, S., et al., *Reciprocally Retained Genes in the Angiosperm Lineage Show the Hallmarks of Dosage Balance Sensitivity.* Plant Cell, 2017.
239. van Hoof, A., *Conserved functions of yeast genes support the duplication, degeneration and complementation model for gene duplication.* Genetics, 2005. **171**(4): p. 1455-61.
240. Conant, G.C., J.a. Birchler, and J.C. Pires, *Dosage, duplication, and diploidization: clarifying the interplay of multiple models for duplicate gene evolution over time.* Current opinion in plant biology, 2014. **19C**: p. 91-98.
241. Schoenfelder, K.P. and D.T. Fox, *The expanding implications of polyploidy.* Journal of Cell Biology, 2015. **209**(4): p. 485-491.
242. Proost, S., et al., *PLAZA 3.0: an access point for plant comparative genomics.* Nucleic Acids Res, 2015. **43**(Database issue): p. D974-81.
243. Hanada, K., et al., *Importance of lineage-specific expansion of plant tandem duplicates in the adaptive response to environmental stimuli.* Plant Physiol, 2008. **148**(2): p. 993-1003.
244. Qian, W. and J. Zhang, *Gene dosage and gene duplicability.* Genetics, 2008. **179**: p. 2319-24.
245. Estep, M.C., et al., *Allopolyploidy, diversification, and the Miocene grassland expansion.* Proc Natl Acad Sci U S A, 2014. **111**(42): p. 15149-54.
246. Vukasinovic, N., et al., *Dissecting a hidden gene duplication: the Arabidopsis thaliana SEC10 locus.* PLoS One, 2014. **9**(4): p. e94077.
247. Rody, H.V., et al., *Both mechanism and age of duplications contribute to biased gene retention patterns in plants.* BMC Genomics, 2017. **18**(1): p. 46.
248. Jubb, H.C., et al., *Mutations at protein-protein interfaces: Small changes over big surfaces have large impacts on human health.* Prog Biophys Mol Biol, 2017. **128**: p. 3-13.
249. Lin, Z. and W.H. Li, *Expansion of hexose transporter genes was associated with the evolution of aerobic fermentation in yeasts.* Mol Biol Evol, 2011. **28**(1): p. 131-42.
250. Sugino, R.P. and H. Innan, *Selection for more of the same product as a force to enhance concerted evolution of duplicated genes.* Trends Genet, 2006. **22**(12): p. 642-4.
251. Proost, S., et al., *PLAZA: a comparative genomics resource to study gene and genome evolution in plants.* Plant Cell, 2009. **21**(12): p. 3718-31.
252. Proost, S., et al., *i-ADHoRe 3.0--fast and sensitive detection of genomic homology in extremely large data sets.* Nucleic Acids Res, 2012. **40**(2): p. e11.
253. Hall, B.G., *Phylogenetic Trees Made Easy.* (Sunderland, MA: Sinauer Associates), 2004.
254. Suyama, M., D. Torrents, and P. Bork, *PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments.* Nucleic Acids Res, 2006. **34**(Web Server issue): p. W609-12.

255. Yang, Z., *PAML 4: phylogenetic analysis by maximum likelihood.* Mol Biol Evol, 2007. **24**(8): p. 1586-91.
256. Van Bel, M. and F. Coppens, *Exploring Plant Co-Expression and Gene-Gene Interactions with CORNET 3.0.* Methods Mol Biol, 2017. **1533**: p. 201-212.
257. Bolger, A.M., M. Lohse, and B. Usadel, *Trimmomatic: a flexible trimmer for Illumina sequence data.* Bioinformatics, 2014. **30**(15): p. 2114-20.
258. Wu, T.D., et al., *GMAP and GSNAP for Genomic Sequence Alignment: Enhancements to Speed, Accuracy, and Functionality.* Methods Mol Biol, 2016. **1418**: p. 283-334.
259. Anders, S., P.T. Pyl, and W. Huber, *HTSeq--a Python framework to work with high-throughput sequencing data.* Bioinformatics, 2015. **31**(2): p. 166-9.
260. Robinson, M.D., D.J. McCarthy, and G.K. Smyth, *edgeR: a Bioconductor package for differential expression analysis of digital gene expression data.* Bioinformatics, 2010. **26**(1): p. 139-40.
261. Chatr-Aryamontri, A., et al., *The BioGRID interaction database: 2013 update.* Nucleic acids research, 2013. **41**: p. D816-23.
262. Jones, A.M., et al., *Border control--a membrane-linked interactome of Arabidopsis.* Science, 2014. **344**(6185): p. 711-6.
263. Bodt, S.D. and J. Hollunder, *CORNET 2.0: integrating plant coexpression, protein–protein interactions, regulatory interactions, gene associations and functional annotations.* New …, 2012: p. 707-720.
264. Van Landeghem, S., et al., *Large-scale event extraction from literature with multi-level gene normalization.* PloS one, 2013. **8**: p. e55814.
265. Cromer, L., et al., *Centromeric cohesion is protected twice at meiosis, by SHUGOSHINs at anaphase I and by PATRONUS at interkinesis.* Curr Biol, 2013. **23**(21): p. 2090-9.
266. Cuellar Perez, A., et al., *The non-JAZ TIFY protein TIFY8 from Arabidopsis thaliana is a transcriptional repressor.* PLoS One, 2014. **9**(1): p. e84891.
267. Liu, S. and C. Trapnell, *Single-cell transcriptome sequencing: recent advances and remaining challenges.* F1000Res, 2016. **5**.
268. Zhao, S. and B. Zhang, *A comprehensive evaluation of ensembl, RefSeq, and UCSC annotations in the context of RNA-seq read mapping and gene quantification.* BMC Genomics, 2015. **16**: p. 97.
269. Hirsch, C.D., N.M. Springer, and C.N. Hirsch, *Genomic limitations to RNA sequencing expression profiling.* Plant J, 2015. **84**(3): p. 491-503.
270. Jaitin, D., I. Amit, and H. KEREN-SHAUL, *High throughput transcriptome analysis*. 2014, Google Patents.
271. Reuter, J.A., D.V. Spacek, and M.P. Snyder, *High-throughput sequencing technologies.* Mol Cell, 2015. **58**(4): p. 586-97.
272. Philippe, N., et al., *Combining DGE and RNA-sequencing data to identify new polyA+ non-coding transcripts in the human genome.* Nucleic Acids Res, 2014. **42**(5): p. 2820-32.
273. Kivioja, T., et al., *Counting absolute numbers of molecules using unique molecular identifiers.* Nat Methods, 2011. **9**(1): p. 72-4.
274. Shen, Y., et al., *Transcriptome dynamics through alternative polyadenylation in developmental and environmental responses in plants revealed by deep sequencing.* Genome Res, 2011. **21**(9): p. 1478-86.
275. Townsley, B.T., et al., *BrAD-seq: Breath Adapter Directional sequencing: a streamlined, ultra-simple and fast library preparation protocol for strand specific mRNA library construction.* Front Plant Sci, 2015. **6**: p. 366.
276. Matz, M., et al., *Amplification of cDNA ends based on template-switching effect and step-out PCR.* Nucleic Acids Res, 1999. **27**(6): p. 1558-60.
277. Force, A., et al., *Preservation of duplicate genes by complementary, degenerative mutations.* Genetics, 1999. **151**(4): p. 1531-45.
278. Van de Peer, Y., et al., *The ghost of selection past: rates of evolution and functional divergence of anciently duplicated genes.* J Mol Evol, 2001. **53**(4-5): p. 436-46.
279. Navratilova, P., et al., *Regulatory divergence of the duplicated chromosomal loci sox11a/b by subpartitioning and sequence evolution of enhancers in zebrafish.* Mol Genet Genomics, 2010. **283**(2): p. 171-84.

280. Bahar Halpern, K., et al., *Single-cell spatial reconstruction reveals global division of labour in the mammalian liver.* Nature, 2017. **542**(7641): p. 352-356.
281. Jaitin, D.A., et al., *Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types.* Science, 2014. **343**(6172): p. 776-9.
282. Lavin, Y., et al., *Tissue-resident macrophage enhancer landscapes are shaped by the local microenvironment.* Cell, 2014. **159**(6): p. 1312-26.
283. Paul, F., et al., *Transcriptional Heterogeneity and Lineage Commitment in Myeloid Progenitors.* Cell, 2015. **163**(7): p. 1663-77.
284. Matcovitch-Natan, O., et al., *Microglia development follows a stepwise program to regulate brain homeostasis.* Science, 2016. **353**(6301): p. aad8670.
285. Jaitin, D.A., et al., *Dissecting Immune Circuits by Linking CRISPR-Pooled Screens with Single-Cell RNA-Seq.* Cell, 2016. **167**(7): p. 1883-1896.e15.
286. Keren-Shaul, H., et al., *A Unique Microglia Type Associated with Restricting Development of Alzheimer's Disease.* Cell, 2017. **169**(7): p. 1276-1290 e17.
287. Bolger, A., et al., *The genome of the stress-tolerant wild tomato species Solanum pennellii.* Nat Genet, 2014. **46**(9): p. 1034-8.
288. Consortium, T.G., *The tomato genome sequence provides insights into fleshy fruit evolution.* Nature, 2012. **485**(7400): p. 635-41.
289. Guillaumet-Adkins, A., et al., *Single-cell transcriptome conservation in cryopreserved cells and tissues.* Genome Biol, 2017. **18**(1): p. 45.
290. Soltis, P.S., et al., *Polyploidy and genome evolution in plants.* Curr Opin Genet Dev, 2015. **35**: p. 119-25.
291. Wendel, J.F., et al., *Evolution of plant genome architecture.* Genome Biol, 2016. **17**: p. 37.
292. Maret, D., et al., *Role of mRNA transcript stability in modulation of expression of the gene encoding thrombin activable fibrinolysis inhibitor.* J Thromb Haemost, 2004. **2**(11): p. 1969-79.
293. Chomczynski, P., *A reagent for the single-step simultaneous isolation of RNA, DNA and proteins from cell and tissue samples.* Biotechniques, 1993. **15**(3): p. 532-4, 536-7.
294. Zhong, S., et al., *High-throughput illumina strand-specific RNA sequencing library preparation.* Cold Spring Harb Protoc, 2011. **2011**(8): p. 940-9.
295. Langmead, B. and S.L. Salzberg, *Fast gapped-read alignment with Bowtie 2.* Nat Methods, 2012. **9**(4): p. 357-9.
296. Trapnell, C., L. Pachter, and S.L. Salzberg, *TopHat: discovering splice junctions with RNA-Seq.* Bioinformatics, 2009. **25**(9): p. 1105-11.
297. Abeel, T., et al., *GenomeView: a next-generation genome browser.* Nucleic Acids Res, 2012. **40**(2): p. e12.
298. Haas, B.J., et al., *Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments.* Genome Biol, 2008. **9**(1): p. R7.
299. Wang, L., et al., *DEGseq: an R package for identifying differentially expressed genes from RNA-seq data.* Bioinformatics (Oxford, England), 2010. **26**: p. 136-8.
300. Spitz, F. and E.E.M. Furlong, *Transcription factors: from enhancer binding to developmental control.* Nature Reviews Genetics, 2012. **13**(9): p. 613-626.
301. Megraw, M., et al., *Small Genetic Circuits and MicroRNAs: Big Players in Polymerase II Transcriptional Control in Plants.* Plant Cell, 2016. **28**(2): p. 286-303.
302. Smith, N.C. and J.M. Matthews, *Mechanisms of DNA-binding specificity and functional gene regulation by transcription factors.* Curr Opin Struct Biol, 2016. **38**: p. 68-74.
303. Dixon, S.J., et al., *Systematic mapping of genetic interaction networks.* Annu Rev Genet, 2009. **43**: p. 601-25.
304. Schmitz, J.F., F. Zimmer, and E. Bornberg-Bauer, *Mechanisms of transcription factor evolution in Metazoa.* Nucleic Acids Res, 2016. **44**(13): p. 6287-97.
305. Zhao, M., et al., *Evolutionary patterns and coevolutionary consequences of MIRNA genes and microRNA targets triggered by multiple mechanisms of genomic duplications in soybean.* Plant Cell, 2015. **27**(3): p. 546-62.

306. Manke, T., R. Bringas, and M. Vingron, *Correlating protein-DNA and protein-protein interaction networks.* J Mol Biol, 2003. **333**(1): p. 75-85.
307. Yu, H., et al., *Genomic analysis of gene expression relationships in transcriptional regulatory networks.* Trends Genet, 2003. **19**(8): p. 422-7.
308. Costanzo, M., et al., *The genetic landscape of a cell.* Science, 2010. **327**(5964): p. 425-31.
309. Martinez, N.J., et al., *A C. elegans genome-scale microRNA network contains composite feedback motifs with high flux capacity.* Genes Dev, 2008. **22**(18): p. 2535-49.
310. Guo, Y., et al., *Integrated network analysis reveals distinct regulatory roles of transcription factors and microRNAs.* RNA, 2016.
311. Mitra, K., et al., *Integrative approaches for finding modular structure in biological networks.* Nat Rev Genet, 2013. **14**(10): p. 719-32.
312. Levine, M. and E.H. Davidson, *Gene regulatory networks for development.* Proc Natl Acad Sci U S A, 2005. **102**(14): p. 4936-42.
313. Milo, R., et al., *Network motifs: simple building blocks of complex networks.* Science, 2002. **298**(5594): p. 824-7.
314. Shen-Orr, S.S., et al., *Network motifs in the transcriptional regulation network of Escherichia coli.* Nat Genet, 2002. **31**(1): p. 64-8.
315. Mangan, S. and U. Alon, *Structure and function of the feed-forward loop network motif.* Proc Natl Acad Sci U S A, 2003. **100**(21): p. 11980-5.
316. Lee, T.I., et al., *Transcriptional regulatory networks in Saccharomyces cerevisiae.* Science, 2002. **298**(5594): p. 799-804.
317. Odom, D.T., et al., *Control of pancreas and liver gene expression by HNF transcription factors.* Science, 2004. **303**(5662): p. 1378-81.
318. Boyer, L.A., et al., *Core Transcriptional Regulatory Circuitry in Human Embryonic Stem Cells.* Cell, 2005. **122**(6): p. 947-956.
319. Tsang, J., J. Zhu, and A. van Oudenaarden, *MicroRNA-mediated feedback and feedforward loops are recurrent network motifs in mammals.* Molecular cell, 2007. **26**: p. 753-67.
320. Cheng, C., et al., *Construction and analysis of an integrated regulatory network derived from high-throughput sequencing data.* PLoS Comput Biol, 2011. **7**(11): p. e1002190.
321. Shalgi, R., et al., *Global and local architecture of the mammalian microRNA-transcription factor regulatory network.* PLoS computational biology, 2007. **3**: p. e131.
322. Ptacek, J., et al., *Global analysis of protein phosphorylation in yeast.* Nature, 2005. **438**(7068): p. 679-684.
323. Zhang, J., et al., *A novel framework for inferring condition-specific TF and miRNA co-regulation of protein-protein interactions.* Gene, 2016. **577**(1): p. 55-64.
324. Atay, O., A. Doncic, and J.M. Skotheim, *Switch-like Transitions Insulate Network Motifs to Modularize Biological Networks.* Cell Syst, 2016.
325. Payne, J.L. and A. Wagner, *Function does not follow form in gene regulatory circuits.* Sci Rep, 2015. **5**: p. 13015.
326. Mazurie, A., S. Bottani, and M. Vergassola, *An evolutionary and functional assessment of regulatory network motifs.* Genome Biol, 2005. **6**(4): p. R35.
327. Dobrin, R., et al., *Aggregation of topological motifs in the Escherichia coli transcriptional regulatory network.* BMC Bioinformatics, 2004. **5**: p. 10.
328. Kashtan, N., et al., *Topological generalizations of network motifs.* Phys Rev E Stat Nonlin Soft Matter Phys, 2004. **70**(3 Pt 1): p. 031909.
329. Zhang, L.V., et al., *Motifs, themes and thematic maps of an integrated Saccharomyces cerevisiae interaction network.* J Biol, 2005. **4**(2): p. 6.
330. Michoel, T., et al., *Enrichment and aggregation of topological motifs are independent organizational principles of integrated interaction networks.* Molecular bioSystems, 2011. **7**: p. 2769-78.
331. Benson, A.R., D.F. Gleich, and J. Leskovec, *Higher-order organization of complex networks.* Science, 2016. **353**(6295): p. 163-166.
332. Tautz, D. and T. Domazet-Loso, *The evolutionary origin of orphan genes.* Nat Rev Genet, 2011. **12**(10): p. 692-702.

333. Domazet-Loso, T., J. Brajkovic, and D. Tautz, *A phylostratigraphy approach to uncover the genomic history of major adaptations in metazoan lineages.* Trends Genet, 2007. **23**(11): p. 533-9.
334. Chen, C.Y., et al., *Dissecting the human protein-protein interaction network via phylogenetic decomposition.* Sci Rep, 2014. **4**: p. 7153.
335. Zhang, W., et al., *New genes drive the evolution of gene interaction networks in the human and mouse genomes.* Genome Biol, 2015. **16**: p. 202.
336. Wei, W., et al., *Genomic Complexity Places Less Restrictions on the Evolution of Young Coexpression Networks than Protein-Protein Interactions.* Genome Biol Evol, 2016. **8**(8): p. 2624-31.
337. Ruprecht, C., et al., *Phylogenomic analysis of gene co-expression networks reveals the evolution of functional modules.* Plant J, 2017. **90**(3): p. 447-465.
338. Capra, J.A., K.S. Pollard, and M. Singh, *Novel genes exhibit distinct patterns of function acquisition and network integration.* Genome Biol, 2010. **11**(12): p. R127.
339. Liu, Z., et al., *Evidence for the additions of clustered interacting nodes during the evolution of protein interaction networks from network motifs.* BMC Evol Biol, 2011. **11**: p. 133.
340. Kim, W.K. and E.M. Marcotte, *Age-Dependent Evolution of the Yeast Protein Interaction Network Suggests a Limited Role of Gene Duplication and Divergence.* PLoS Comput Biol, 2008. **4**(11): p. e1000232.
341. Liang, C., J. Luo, and D. Song, *Network simulation reveals significant contribution of network motifs to the age-dependency of yeast protein-protein interaction networks.* Mol Biosyst, 2014. **10**(9): p. 2277-88.
342. Lipinski, K.J., et al., *High spontaneous rate of gene duplication in Caenorhabditis elegans.* Curr Biol, 2011. **21**(4): p. 306-10.
343. Arsovski, A.A., et al., *Evolution of cis-regulatory elements and regulatory networks in duplicated genes of Arabidopsis thaliana.* Plant Physiology, 2015. **169**: p. pp.00717.2015.
344. Barabasi, A.L. and Z.N. Oltvai, *Network biology: understanding the cell's functional organization.* Nat Rev Genet, 2004. **5**(2): p. 101-13.
345. Newman, M.E.J., *The structure and function of complex networks.* Society for Industrial and Applied Mathematics Review, 2003. **45**(2): p. 167-256.
346. Demeyer, S., et al., *The index-based subgraph matching algorithm (ISMA): fast subgraph enumeration in large networks using optimized search trees.* PLoS One, 2013. **8**(4): p. e61183.
347. Reece-Hoyes, J.S., et al., *Extensive rewiring and complex evolutionary dynamics in a C. elegans multiparameter transcription factor network.* Mol Cell, 2013. **51**(1): p. 116-27.
348. Gerstein, M.B., et al., *Architecture of the human regulatory network derived from ENCODE data.* Nature, 2012. **489**(7414): p. 91-100.
349. Vermeirssen, V., et al., *Transcription factor modularity in a gene-centered C. elegans core neuronal protein-DNA interaction network.* Genome Research, 2007. **17**(7): p. 1061-71.
350. Michoel, T. and B. Nachtergaele, *Alignment and integration of complex networks by hypergraph-based spectral clustering.* Phys Rev E Stat Nonlin Soft Matter Phys, 2012. **86**(5 Pt 2): p. 056111.
351. Levin, M., et al., *Developmental Milestones Punctuate Gene Expression in the Caenorhabditis Embryo.* Developmental Cell, 2012. **22**: p. 1101-1108.
352. Spencer, W.C., et al., *A spatial and temporal map of C. elegans gene expression.* Genome research, 2011. **21**: p. 325-41.
353. Vermeirssen, V., et al., *Arabidopsis ensemble reverse-engineered gene regulatory network discloses interconnected transcription factors in oxidative stress.* Plant Cell, 2014. **26**(12): p. 4656-79.
354. Kelley, R. and T. Ideker, *Systematic interpretation of genetic interactions using protein networks.* Nat Biotechnol, 2005. **23**(5): p. 561-6.
355. Nishikori, S., et al., *p97 Homologs from Caenorhabditis elegans, CDC-48.1 and CDC-48.2, suppress the aggregate formation of huntingtin exon1 containing expanded polyQ repeat.* Genes Cells, 2008. **13**(8): p. 827-38.
356. Furuya, M., et al., *The C. elegans eyes absent ortholog EYA-1 is required for tissue differentiation and plays partially redundant roles with PAX-6.* Dev Biol, 2005. **286**(2): p. 452-63.

357. Nag, A., S. King, and T. Jack, *miR319a targeting of TCP4 is critical for petal growth and development in Arabidopsis.* Proceedings of the National Academy of Sciences of the United States of America, 2009. **106**(52): p. 22534-22539.
358. Wang, H. and H. Wang, *The miR156/SPL Module, a Regulatory Hub and Versatile Toolbox, Gears up Crops for Enhanced Agronomic Traits.* Molecular Plant, 2015. **8**(5): p. 677-688.
359. Hwan Lee, J., J. Joon Kim, and J.H. Ahn, *Role of SEPALLATA3 (SEP3) as a downstream gene of miR156-SPL3-FT circuitry in ambient temperature-responsive flowering.* Plant Signal Behav, 2012. **7**(9): p. 1151-4.
360. Tan, Q.K. and V.F. Irish, *The Arabidopsis zinc finger-homeodomain genes encode proteins with unique biochemical properties that are coordinately expressed during floral development.* Plant Physiol, 2006. **140**(3): p. 1095-108.
361. Tran, L.S., et al., *Co-expression of the stress-inducible zinc finger homeodomain ZFHD1 and NAC transcription factors enhances expression of the ERD1 gene in Arabidopsis.* Plant J, 2007. **49**(1): p. 46-63.
362. Wang, W., et al., *Genome-wide analysis and expression patterns of ZF-HD transcription factors under different developmental tissues and abiotic stresses in Chinese cabbage.* Mol Genet Genomics, 2016. **291**(3): p. 1451-64.
363. Wang, H., et al., *Genome-wide identification, evolution and expression analysis of the grape (Vitis vinifera L.) zinc finger-homeodomain gene family.* Int J Mol Sci, 2014. **15**(4): p. 5730-48.
364. Saez, A., et al., *HAB1-SWI3B interaction reveals a link between abscisic acid signaling and putative SWI/SNF chromatin-remodeling complexes in Arabidopsis.* Plant Cell, 2008. **20**(11): p. 2972-88.
365. Sarnowski, T.J., et al., *SWI3 subunits of putative SWI/SNF chromatin-remodeling complexes play distinct roles during Arabidopsis development.* Plant Cell, 2005. **17**(9): p. 2454-72.
366. Ko, J.H., et al., *The MYB46/MYB83-mediated transcriptional regulatory programme is a gatekeeper of secondary wall biosynthesis.* Ann Bot, 2014. **114**(6): p. 1099-107.
367. Haake, V., et al., *Transcription factor CBF4 is a regulator of drought adaptation in Arabidopsis.* Plant Physiol, 2002. **130**(2): p. 639-48.
368. Shaikhali, J., et al., *The CRYPTOCHROME1-dependent response to excess light is mediated through the transcriptional activators ZINC FINGER PROTEIN EXPRESSED IN INFLORESCENCE MERISTEM LIKE1 and ZML2 in Arabidopsis.* Plant Cell, 2012. **24**(7): p. 3009-25.
369. McFarlane, H.E., A. Doring, and S. Persson, *The cell biology of cellulose synthesis.* Annu Rev Plant Biol, 2014. **65**: p. 69-94.
370. Vain, T., et al., *The Cellulase KORRIGAN Is Part of the Cellulose Synthase Complex.* Plant Physiol, 2014. **165**(4): p. 1521-1532.
371. Xie, L., C. Yang, and X. Wang, *Brassinosteroids can regulate cellulose biosynthesis by controlling the expression of CESA genes in Arabidopsis.* Journal of Experimental Botany, 2011. **62**(13): p. 4495-4506.
372. Guo, H., et al., *Expression of the MYB transcription factor gene BplMYB46 affects abiotic stress tolerance and secondary cell wall deposition in Betula platyphylla.* Plant Biotechnol J, 2017. **15**(1): p. 107-121.
373. Endler, A., et al., *A Mechanism for Sustained Cellulose Synthesis during Salt Stress.* Cell, 2015. **162**(6): p. 1353-64.
374. Li, Y., et al., *Novel insights into the function of Arabidopsis R2R3-MYB transcription factors regulating aliphatic glucosinolate biosynthesis.* Plant Cell Physiol, 2013. **54**(8): p. 1335-44.
375. Frerigmann, H. and T. Gigolashvili, *MYB34, MYB51, and MYB122 distinctly regulate indolic glucosinolate biosynthesis in Arabidopsis thaliana.* Mol Plant, 2014. **7**(5): p. 814-28.
376. Li, Y., et al., *Expansion of biological pathways based on evolutionary inference.* Cell, 2014. **158**: p. 213-25.
377. Martinez-Ballesta, M., et al., *The impact of the absence of aliphatic glucosinolates on water transport under salt stress in Arabidopsis thaliana.* Front Plant Sci, 2015. **6**: p. 524.
378. Li, B., et al., *Promoter-based integration in plant defense regulation.* Plant Physiol, 2014. **166**(4): p. 1803-20.
379. Mithen, R., R. Bennett, and J. Marquez, *Glucosinolate biochemical diversity and innovation in the Brassicales.* Phytochemistry, 2010. **71**(17-18): p. 2074-86.

380. Beyer, A., et al., *Integrated assessment and prediction of transcription factor binding.* PLoS Comput Biol, 2006. **2**(6): p. e70.
381. Park, C.Y., et al., *Simultaneous genome-wide inference of physical, genetic, regulatory, and functional pathway components.* PLoS Comput Biol, 2010. **6**(11): p. e1001009.
382. Lee, I., et al., *A probabilistic functional network of yeast genes.* Science, 2004. **306**(5701): p. 1555-8.
383. Zhao, H., et al., *'Traffic light rules': Chromatin states direct miRNA-mediated network motifs running by integrating epigenome and regulatome.* Biochimica et Biophysica Acta (BBA) - General Subjects, 2016. **1860**(7): p. 1475-1488.
384. Beber, M.E., et al., *Artefacts in statistical analyses of network motifs: general framework and application to metabolic networks.* Journal of The Royal Society Interface, 2012. **9**(77): p. 3426-3435.
385. Ginoza, R. and A. Mugler, *Network motifs come in sets: Correlations in the randomization process.* Physical Review E, 2010. **82**(1): p. 011921.
386. Megraw, M., S. Mukherjee, and U. Ohler, *Sustained-input switches for transcription factors and microRNAs are central building blocks of eukaryotic gene circuits.* Genome Biology, 2013. **14**(8): p. R85.
387. Zinman, G.E., S. Zhong, and Z. Bar-Joseph, *Biological interaction networks are conserved at the module level.* BMC systems biology, 2011. **5**: p. 134.
388. Abrusan, G., *Integration of new genes into cellular networks, and their structural maturation.* Genetics, 2013. **195**(4): p. 1407-17.
389. Barabasi, A.L. and R. Albert, *Emergence of scaling in random networks.* Science, 1999. **286**(5439): p. 509-12.
390. Ispolatov, I., P.L. Krapivsky, and A. Yuryev, *Duplication-divergence model of protein interaction network.* Phys Rev E Stat Nonlin Soft Matter Phys, 2005. **71**(6 Pt 1): p. 061911.
391. Abdelzaher, A.F., et al., *Transcriptional Network Growing Models Using Motif-Based Preferential Attachment.* Front Bioeng Biotechnol, 2015. **3**: p. 157.
392. Lehti-Shiu, M.D., et al., *Diversity, expansion, and evolutionary novelty of plant DNA-binding transcription factor families.* Biochim Biophys Acta, 2017. **1860**(1): p. 3-20.
393. Haerty, W., et al., *Comparative analysis of function and interaction of transcription factors in nematodes: extensive conservation of orthology coupled to rapid sequence evolution.* BMC genomics, 2008. **9**: p. 399.
394. Ward, J.J. and J.M. Thornton, *Evolutionary models for formation of network motifs and modularity in the Saccharomyces transcription factor network.* PLoS computational biology, 2007. **3**: p. 1993-2002.
395. Jin, J., et al., *An Arabidopsis Transcriptional Regulatory Map Reveals Distinct Functional and Evolutionary Features of Novel Transcription Factors.* Molecular Biology and Evolution, 2015.
396. Harris, T.W., et al., *WormBase 2014: new views of curated biology.* Nucleic Acids Res, 2014. **42**(Database issue): p. D789-93.
397. Simonis, N., et al., *Empirically controlled mapping of the Caenorhabditis elegans protein-protein interactome network.* Nat Methods, 2009. **6**(1): p. 47-54.
398. Chatr-Aryamontri, A., et al., *The BioGRID interaction database: 2015 update.* Nucleic Acids Res, 2015. **43**(Database issue): p. D470-8.
399. Cheeseman, I.M., et al., *The conserved KMN network constitutes the core microtubule-binding site of the kinetochore.* Cell, 2006. **127**(5): p. 983-97.
400. Cheeseman, I.M., et al., *A conserved protein network controls assembly of the outer kinetochore and its ability to sustain tension.* Genes Dev, 2004. **18**(18): p. 2255-68.
401. Grove, C.A., et al., *A multiparameter network reveals extensive divergence between C. elegans bHLH transcription factors.* Cell, 2009. **138**(2): p. 314-27.
402. Popovici, C., et al., *Direct and heterologous approaches to identify the LET-756/FGF interactome.* BMC Genomics, 2006. **7**: p. 105.
403. Byrne, A.B., et al., *A global analysis of genetic interactions in Caenorhabditis elegans.* J Biol, 2007. **6**(3): p. 8.
404. Lehner, B., et al., *Systematic mapping of genetic interactions in Caenorhabditis elegans identifies common modifiers of diverse signaling pathways.* Nat Genet, 2006. **38**(8): p. 896-903.

405. Tischler, J., et al., *Combinatorial RNA interference in Caenorhabditis elegans reveals that redundancy between gene duplicates can be maintained for more than 80 million years of evolution.* Genome Biol, 2006. **7**(8): p. R69.
406. Tischler, J., B. Lehner, and A.G. Fraser, *Evolutionary plasticity of genetic interaction networks.* Nat Genet, 2008. **40**(4): p. 390-1.
407. Reinke, V., M. Krause, and P. Okkema, *Transcriptional regulation of gene expression in C. elegans.* WormBook, 2013: p. 1-34.
408. Kozomara, A. and S. Griffiths-Jones, *miRBase: annotating high confidence microRNAs using deep sequencing data.* Nucleic Acids Research, 2014. **42**(D1): p. D68-D73.
409. Arda, H.E., et al., *Functional modularity of nuclear hormone receptors in a Caenorhabditis elegans metabolic gene regulatory network.* Mol Syst Biol, 2010. **6**: p. 367.
410. Deplancke, B., et al., *A gateway-compatible yeast one-hybrid system.* Genome Res, 2004. **14**(10B): p. 2093-101.
411. Deplancke, B., et al., *A gene-centered C. elegans protein-DNA interaction network.* Cell, 2006. **125**(6): p. 1193-205.
412. Feng, H., et al., *A regulatory cascade of three transcription factors in a single specific neuron, DVC, in Caenorhabditis elegans.* Gene, 2012. **494**(1): p. 73-84.
413. Reece-Hoyes, J.S., et al., *The C. elegans Snail homolog CES-1 can activate gene expression in vivo and share targets with bHLH transcription factors.* Nucleic Acids Res, 2009. **37**(11): p. 3689-98.
414. Reece-Hoyes, J.S., et al., *Enhanced yeast one-hybrid assays for high-throughput gene-centered regulatory network mapping.* Nat Methods, 2011. **8**(12): p. 1059-64.
415. Vermeirssen, V., et al., *Matrix and Steiner-triple-system smart pooling assays for high-performance transcription regulatory network mapping.* Nat Methods, 2007. **4**(8): p. 659-64.
416. Boyle, A.P., et al., *Comparative analysis of regulatory information and circuits across distant species.* Nature, 2014. **512**(7515): p. 453-456.
417. Kirienko, N.V. and D.S. Fay, *Transcriptome profiling of the C. elegans Rb ortholog reveals diverse developmental roles.* Dev Biol, 2007. **305**(2): p. 674-84.
418. Kouns, N.A., et al., *NHR-23 dependent collagen and hedgehog-related genes required for molting.* Biochem Biophys Res Commun, 2011. **413**(4): p. 515-20.
419. Magner, D.B., et al., *The NHR-8 nuclear receptor regulates cholesterol and bile acid homeostasis in C. elegans.* Cell Metab, 2013. **18**(2): p. 212-24.
420. Pathare, P.P., et al., *Coordinate regulation of lipid metabolism by novel nuclear receptor partnerships.* PLoS Genet, 2012. **8**(4): p. e1002645.
421. Petrella, L.N., et al., *synMuv B proteins antagonize germline fate in the intestine and ensure C. elegans survival.* Development, 2011. **138**(6): p. 1069-79.
422. Thyagarajan, B., et al., *ETS-4 is a transcriptional regulator of life span in Caenorhabditis elegans.* PLoS Genet, 2010. **6**(9): p. e1001125.
423. Troemel, E.R., et al., *p38 MAPK regulates expression of immune response genes and contributes to longevity in C. elegans.* PLoS Genet, 2006. **2**(11): p. e183.
424. Van Nostrand, E.L. and S.K. Kim, *Integrative analysis of C. elegans modENCODE ChIP-seq datasets to infer transcription factor-responsive targets and upstream regulators of differentially-expressed genes from expression profiling experiments.* Genome Research, 2013.
425. Hsu, S.D., et al., *miRTarBase update 2014: an information resource for experimentally validated miRNA-target interactions.* Nucleic Acids Res, 2014. **42**(Database issue): p. D78-85.
426. Gerstein, M.B., et al., *Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project.* Science (New York, N.Y.), 2010. **330**: p. 1775-87.
427. Lall, S., et al., *A genome-wide map of conserved microRNA targets in C. elegans.* Curr Biol, 2006. **16**(5): p. 460-71.
428. Engelmann, I., et al., *A comprehensive analysis of gene expression changes provoked by bacterial and fungal infection in C. elegans.* PLoS One, 2011. **6**(5): p. e19055.
429. Spieth, J., et al., *Overview of gene structure in C. elegans.* WormBook, 2014: p. 1-18.
430. Jin, J., et al., *PlantTFDB 3.0: a portal for the functional and evolutionary study of plant transcription factors.* Nucleic Acids Res, 2014. **42**(Database issue): p. D1182-7.

431. Heyndrickx, K.S., et al., *A functional and evolutionary perspective on transcription factor binding in Arabidopsis thaliana.* Plant Cell, 2014. **26**(10): p. 3894-910.
432. Brady, S.M., et al., *A stele-enriched gene regulatory network in the Arabidopsis root.* Molecular Systems Biology, 2011. **7**: p. 459.
433. Taylor-Teeples, M., et al., *An Arabidopsis gene regulatory network for secondary cell wall synthesis.* Nature, 2015. **517**(7536): p. 571-5.
434. Yilmaz, A., et al., *AGRIS: the Arabidopsis Gene Regulatory Information Server, an update.* Nucleic acids research, 2011. **39**: p. D1118-22.
435. Srivastava, P.K., et al., *A comparison of performance of plant miRNA target prediction tools and the characterization of features for genome-wide target prediction.* BMC Genomics, 2014. **15**(1): p. 1-15.
436. Berardini, T.Z., et al., *The Arabidopsis information resource: Making and mining the "gold standard" annotated reference plant genome.* Genesis, 2015. **53**(8): p. 474-85.
437. Csardi, G. and T. Nepusz, *The igraph software package for complex network research.* InterJournal, Complex Systems, 2006. **1695**(5): p. 1-9.
438. Chen, L., et al., *Identification of breast cancer patients based on human signaling network motifs.* Scientific Reports, 2013. **3**: p. 3368.
439. Liebeskind, B.J., C.D. McWhite, and E.M. Marcotte, *Towards Consensus Gene Ages.* Genome Biol Evol, 2016. **8**(6): p. 1812-23.
440. Zhou, K., et al., *Genome-wide identification of lineage-specific genes within Caenorhabditis elegans.* Genomics, 2015. **106**(4): p. 242-8.
441. Heyndrickx, K.S. and K. Vandepoele, *Systematic identification of functional plant modules through the integration of complementary data sources.* Plant physiology, 2012. **159**: p. 884-901.
442. Birchler, J.A. and R.A. Veitia, *Gene balance hypothesis: connecting issues of dosage sensitivity across biological disciplines.* Proceedings of the National Academy of Sciences of the United States of America, 2012. **109**: p. 14746-53.
443. Acharya, D. and T.C. Ghosh, *Global analysis of human duplicated genes reveals the relative importance of whole-genome duplicates originated in the early vertebrate evolution.* Bmc Genomics, 2016. **17**.
444. Ha, M., W.H. Li, and Z.J. Chen, *External factors accelerate expression divergence between duplicate genes.* Trends Genet, 2007. **23**(4): p. 162-6.
445. Soltis, P.S. and D.E. Soltis, *Ancient WGD events as drivers of key innovations in angiosperms.* Curr Opin Plant Biol, 2016. **30**: p. 159-65.
446. Brochmann, C., et al., *Polyploidy in arctic plants.* Biological Journal of the Linnean Society, 2004. **82**(4): p. 521-536.
447. Ramsey, J., *Polyploidy and ecological adaptation in wild yarrow.* Proceedings of the National Academy of Sciences of the United States of America, 2011. **108**(17): p. 7096-7101.
448. Freeling, M., *Picking up the Ball at the K/Pg Boundary: The Distribution of Ancient Polyploidies in the Plant Phylogenetic Tree as a Spandrel of Asexuality with Occasional Sex.* Plant Cell, 2017. **29**(2): p. 202-206.
449. Lohaus, R. and Y. Van de Peer, *Of dups and dinos: evolution at the K/Pg boundary.* Curr Opin Plant Biol, 2016. **30**: p. 62-9.
450. McLysaght, A., et al., *Ohnologs are overrepresented in pathogenic copy number mutations.* Proceedings of the National Academy of Sciences of the United States of America, 2014. **111**: p. 361-6.
451. Singh, P.P., et al., *On the Expansion of "Dangerous" Gene Repertoires by Whole-Genome Duplications in Early Vertebrates.* Cell Reports, 2012. **2**(5): p. 1387-1398.
452. Rice, A.M. and A. McLysaght, *Dosage sensitivity is a major determinant of human copy number variant pathogenicity.* Nat Commun, 2017. **8**: p. 14366.
453. Rice, A.M. and A. McLysaght, *Dosage-sensitive genes in evolution and disease.* BMC Biol, 2017. **15**(1): p. 78.
454. Conrad, B. and S.E. Antonarakis, *Gene duplication: a drive for phenotypic diversity and cause of human disease.* Annu Rev Genomics Hum Genet, 2007. **8**: p. 17-35.

455. Deane, C.M., et al., *Protein interactions: two methods for assessment of the reliability of high throughput observations.* Mol Cell Proteomics, 2002. **1**(5): p. 349-56.
456. Szklarczyk, D., et al., *STRING v10: protein-protein interaction networks, integrated over the tree of life.* Nucleic Acids Res, 2015. **43**(Database issue): p. D447-52.
457. Pereira-Leal, J.B., et al., *Evolution of protein complexes by duplication of homomeric interactions.* Genome biology, 2007. **8**: p. R51.
458. Pereira-Leal, J.B., E.D. Levy, and S.a. Teichmann, *The origins and evolution of functional modules: lessons from protein complexes.* Philosophical transactions of the Royal Society of London. Series B, Biological sciences, 2006. **361**: p. 507-517.
459. Pereira-Leal, J.B. and S.a. Teichmann, *Novel specificities emerge by stepwise duplication of functional modules.* Genome Research, 2005. **15**: p. 552-559.
460. Oberdorf, R. and T. Kortemme, *Complex topology rather than complex membership is a determinant of protein dosage sensitivity.* Molecular systems biology, 2009. **5**: p. 253.
461. Kolar, F., et al., *Mixed-Ploidy Species: Progress and Opportunities in Polyploid Research.* Trends Plant Sci, 2017. **22**(12): p. 1041-1055.
462. Certner, M., et al., *Evolutionary dynamics of mixed-ploidy populations in an annual herb: dispersal, local persistence and recurrent origins of polyploids.* Ann Bot, 2017. **120**(2): p. 303-315.
463. Gutierrez, J. and S. Maere, *Modeling the evolution of molecular systems from a mechanistic perspective.* Trends Plant Sci, 2014. **19**(5): p. 292-303.
464. Yang, L., et al., *Lowly expressed genes in Arabidopsis thaliana bear the signature of possible pseudogenization by promoter degradation.* Molecular biology and evolution, 2011.
465. D'Antonio, M. and F.D. Ciccarelli, *Modification of gene duplicability during the evolution of protein interaction network.* PLoS computational biology, 2011. **7**: p. e1002029.
466. Chaudhary, B., et al., *Reciprocal silencing, transcriptional bias and functional divergence of homeologs in polyploid cotton (gossypium).* Genetics, 2009. **182**(2): p. 503-17.
467. Dong, D., Z. Yuan, and Z. Zhang, *Evidences for increased expression variation of duplicate genes in budding yeast: from cis- to trans-regulation effects.* Nucleic Acids Res, 2011. **39**(3): p. 837-47.
468. Leach, L.J., et al., *The role of cis-regulatory motifs and genetical control of expression in the divergence of yeast duplicate genes.* Mol Biol Evol, 2007. **24**(11): p. 2556-65.
469. Zou, C., et al., *Evolution of stress-regulated gene expression in duplicate genes of Arabidopsis thaliana.* PLoS Genet, 2009. **5**(7): p. e1000581.
470. Keller, T.E. and S.V. Yi, *DNA methylation and evolution of duplicate genes.* Proc Natl Acad Sci U S A, 2014. **111**(16): p. 5932-7.
471. Wang, L., et al., *Comparative epigenomics reveals evolution of duplicated genes in potato and tomato.* Plant J, 2017.
472. Cheng, F., et al., *Biased gene fractionation and dominant gene expression among the subgenomes of Brassica rapa.* PloS one, 2012. **7**: p. e36442.
473. Sui, Y., et al., *Genomic, regulatory and epigenetic mechanisms underlying duplicated gene evolution in the natural allotetraploid Oryza minuta.* BMC Genomics, 2014. **15**: p. 11.
474. Iniguez, L.P. and G. Hernandez, *The Evolutionary Relationship between Alternative Splicing and Gene Duplication.* Front Genet, 2017. **8**: p. 14.
475. Kopelman, N.M., D. Lancet, and I. Yanai, *Alternative splicing and gene duplication are inversely correlated evolutionary mechanisms.* Nat Genet, 2005. **37**(6): p. 588-9.
476. Abascal, F., M.L. Tress, and A. Valencia, *The evolutionary fate of alternatively spliced homologous exons after gene duplication.* Genome Biol Evol, 2015. **7**(6): p. 1392-403.
477. Roux, J. and M. Robinson-Rechavi, *Age-dependent gain of alternative splice forms and biased duplication explain the relation between splicing and duplication.* Genome Res, 2011. **21**(3): p. 357-63.
478. Lin, H., et al., *Characterization of paralogous protein families in rice.* BMC Plant Biol, 2008. **8**: p. 18.
479. Tzfadia, O., et al., *The MORPH algorithm: ranking candidate genes for membership in Arabidopsis and tomato pathways.* Plant Cell, 2012. **24**(11): p. 4389-406.
480. Koltai, H. and C. Weingarten-Baror, *Specificity of DNA microarray hybridization: characterization, effectors and approaches for data correction.* Nucleic Acids Res, 2008. **36**(7): p. 2395-405.

481. Jiang, S.Y., J.M. Gonzalez, and S. Ramachandran, *Comparative genomic and transcriptomic analysis of tandemly and segmentally duplicated genes in rice.* PLoS One, 2013. **8**(5): p. e63551.
482. Shafee T, L.R., *Eukaryotic and prokaryotic gene structure.* WikiJournal of Medicine 4 2017.
483. Ulitsky, I., *Evolution to the rescue: using comparative genomics to understand long non-coding RNAs.* Nat Rev Genet, 2016. **17**(10): p. 601-14.
484. Pantano, L., et al., *The small RNA content of human sperm reveals pseudogene-derived piRNAs complementary to protein-coding genes.* RNA, 2015. **21**(6): p. 1085-95.
485. Guo, X., et al., *Characterization of human pseudogene-derived non-coding RNAs for functional potential.* PLoS One, 2014. **9**(4): p. e93972.
486. Poliseno, L., *Pseudogenes: newly discovered players in human cancer.* Sci Signal, 2012. **5**(242): p. re5.
487. Thibaud-Nissen, F., S. Ouyang, and C.R. Buell, *Identification and characterization of pseudogenes in the rice gene complement.* BMC Genomics, 2009. **10**: p. 317.
488. Zou, C., et al., *Evolutionary and expression signatures of pseudogenes in Arabidopsis and rice.* Plant Physiol, 2009. **151**(1): p. 3-15.
489. Wicker, T., et al., *Frequent gene movement and pseudogene evolution is common to the large and complex genomes of wheat, barley, and their relatives.* Plant Cell, 2011. **23**(5): p. 1706-18.
490. Xiao, J., et al., *Pseudogenes and Their Genome-Wide Prediction in Plants.* Int J Mol Sci, 2016. **17**(12).
491. van Baren, M.J. and M.R. Brent, *Iterative gene prediction and pseudogene removal improves genome annotation.* Genome Res, 2006. **16**(5): p. 678-85.
492. Solovyev, V., et al., *Automatic annotation of eukaryotic genes, pseudogenes and promoters.* Genome Biol, 2006. **7 Suppl 1**: p. S10 1-12.
493. Zhang, Z., et al., *PseudoPipe: an automated pseudogene identification pipeline.* Bioinformatics, 2006. **22**(12): p. 1437-9.
494. Geertz, M. and S.J. Maerkl, *Experimental strategies for studying transcription factor-DNA binding specificities.* Brief Funct Genomics, 2010. **9**(5-6): p. 362-73.
495. Van de Velde, J., et al., *A Collection of Conserved Noncoding Sequences to Study Gene Regulation in Flowering Plants.* Plant Physiol, 2016. **171**(4): p. 2586-98.
496. De Witte, D., et al., *BLSSpeller: exhaustive comparative discovery of conserved cis-regulatory elements.* Bioinformatics, 2015. **31**(23): p. 3758-66.